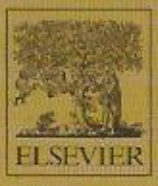


SONIA VIEIRA

INTRODUÇÃO À BIOESTATÍSTICA

3ª Edição

INTRODUÇÃO À BIOESTATÍSTICA

Este livro é instrumento valioso para os profissionais das áreas de saúde que precisam interpretar estatísticas e é obrigatório para aqueles que fazem ou lêem pesquisas científicas.

Diferentemente dos textos que pressupõem conhecimentos avançados na área de matemática, este livro foi escrito para aqueles que pretendem aprender Bioestatística, mas que não são profissionais nessa área.

Com linguagem clara e acessível, o livro apresenta conceitos de forma didática, sempre com muitos exemplos e exercícios. Ainda, o livro cobre grande variedade de assuntos e tem grande flexibilidade. Pode, portanto, ser indicado como texto na disciplina de Bioestatística, tanto a nível de Graduação como Pós-graduação.

SONIA VIEIRA

é professora titular de Bioestatística na UNICAMP. Já esteve, como professora convidada, na Universidade da Califórnia e na Universidade Yale. Além de duas teses acadêmicas e diversos artigos publicados em revistas nacionais e internacionais, escreveu os seguintes livros: *Metodologia científica* (Sarvier), *Como escrever uma Tese* (Pioneira) e, em co-autoria, *Análise e regressão* (Hucitec), *Estatística – Introdução ilustrada* (Atlas), *Elementos de Estatística* (Atlas), *O que é Estatística* (Brasiliense), *Experimentação com seres humanos* (Moderna), *Estatística experimental* (Atlas) e *As 7 ferramentas estatísticas para o controle da qualidade* (QA & T0).

ISBN 13 - 978-85-352-0259-5
ISBN 10 - 85-352-0259-5

Uma empresa Elsevier
www.campus.com.br

SONIA VIEIRA

INTRODUÇÃO À BIOESTATÍSTICA

16ª Tiragem

ELSEVIER

EDITORA CAMPUS

Capa:
Otávio Studart

Copidesque
Jorge Uranga

Editoração Eletrônica
RioTexto

Revisão Gráfica
Gypsi Canetti
Maria da Penha

Projeto Gráfico
Elsevier Editora Ltda.
A Qualidade da Informação.
Rua Sete de Setembro, 111 – 16º andar
20050-006 Rio de Janeiro RJ Brasil
Telefone: (21) 3970-9300 FAX: (21) 2507-1991
E-mail: *info@elsevier.com.br*
Escritório São Paulo:
Rua Quintana, 753/8º andar
04569-011 Brooklin São Paulo SP
Tel.: (11) 5105-8555

ISBN 13: 978-85-352-0259-5
ISBN 10: 85-352-0259-5

CIP-Brasil. Catalogação-na-fonte.
Sindicato Nacional dos Editores de Livros, RJ

Vieira, Sonia, 1942-
V718i Introdução à bioestatística / Sonia Vieira. – 3 ed. revista e
3. ed. ampliada. – Rio de Janeiro : Elsevier, 1980 – 16ª Reimpressão.

Inclui bibliografia e anexos
ISBN 85-352-0259-5

1. Bioestatística. I. Título.

97-1809. CDD – 574.072
CDU – 574.001.5

Apresentação

A ciência não é um conhecimento definitivo sobre a realidade, mas um conhecimento hipotético, que pode ser questionado e corrigido. Ensinar ciência não significa apenas descrever fatos, enunciar leis e apresentar novas descobertas — mas ensinar o método científico, que é a maneira crítica de buscar o conhecimento.

O método científico exige, porém, organizar dados, analisar e tomar decisões em condições de incerteza. Dá suporte técnico a esse trabalho a Estatística, que pode ser vista, pelo pesquisador, como uma ferramenta do método científico. Bioestatística é a Estatística aplicada às ciências médica e biológica.

Pode parecer difícil ao aluno que não tem gosto pela matemática aprender Bioestatística. Mas mesmo o estudante das ciências médica e biológica deve adquirir algum conhecimento desta matéria, pois só assim terá um ponto de vista objetivo sobre as técnicas do método científico e saberá avaliar o grau de importância da informação fornecida por essas técnicas.

Outra conseqüência importante de aprender Bioestatística — mais importante do que possa parecer à primeira vista — é a familiarização com o jargão próprio da área. Alguns termos do vocabulário comum têm significado técnico e específico, quando usados em Bioestatística. É claro que o conhecimento do significado comum ajuda, mas pode conduzir à interpretação errada quando substitui o significado técnico.

Finalmente, ensinar Bioestatística é um desafio, porque a disciplina não pertence ao elenco de disciplinas profissionalizantes dos cursos em que é ministrada, e enfrenta, por isso, o descaso de boa parte dos alunos. Mas este livro foi escrito e reescrito diversas vezes, na tentativa de facilitar a aprendizagem. Os conceitos são transmitidos mais pela intuição do que por demonstração, os exemplos são simples e das áreas médica e biológica,

os exercícios exigem pouco trabalho de cálculo. Enfim — sem despender muito tempo com cálculos e demonstrações — o estudante adquire, neste livro, conhecimentos suficientes para tornar-se usuário competente das técnicas estatísticas mais comuns.

Sumário

CAPÍTULO 1
NOÇÕES BÁSICAS, 1

1.1 - VARIÁVEIS, 1
1.2 - APURAÇÃO DE DADOS, 2
1.3 - POPULAÇÃO E AMOSTRA, 2
1.4 - TÉCNICAS DE AMOSTRAGEM, 3
1.4.1 - Amostra casual simples, 3
1.4.2 - Amostra sistemática, 4
1.4.3 - Amostra estratificada, 5
1.4.4 - Amostra de conveniência, 5
1.5 - EXERCÍCIOS RESOLVIDOS, 6
1.6 - EXERCÍCIOS PROPOSTOS, 7

CAPÍTULO 2
APRESENTAÇÃO DE DADOS EM TABELAS, 9

2.1 - COMPONENTES DAS TABELAS, 9
2.2 - TABELAS DE CONTINGÊNCIA, 11
2.3 - TABELAS DE DISTRIBUIÇÃO DE FREQÜÊNCIAS, 12
2.4 - EXERCÍCIOS RESOLVIDOS, 16
2.5 - EXERCÍCIOS PROPOSTOS, 17

CAPÍTULO 3
APRESENTAÇÃO DE DADOS EM GRÁFICOS, 19

3.1 - GRÁFICO DE BARRAS, 19
3.2 - GRÁFICO DE SETORES, 20
3.3 - HISTOGRAMA, 21
3.4 - POLÍGONO DE FREQÜÊNCIAS, 22
3.5 - EXERCÍCIOS RESOLVIDOS, 24
3.6 - EXERCÍCIOS PROPOSTOS, 26

CAPÍTULO 4
MEDIDAS DE TENDÊNCIA CENTRAL PARA UMA AMOSTRA, 27

4.1 - MÉDIA ARITMÉTICA, 27
4.2 - MÉDIA DE DADOS EM TABELAS DE DISTRIBUIÇÃO DE FREQÜÊNCIAS, 28
4.3 - MEDIANA, 30
4.4 - MODA, 31
4.5 - EXERCÍCIOS RESOLVIDOS, 32
4.6 - EXERCÍCIOS PROPOSTOS, 33

CAPÍTULO 5
MEDIDAS DE DISPERSÃO PARA UMA AMOSTRA, 35

5.1 - AMPLITUDE, 36
5.2 - VARIÂNCIA, 36
5.3 - DESVIO PADRÃO, 40
5.4 - COEFICIENTE DE VARIAÇÃO, 40
5.5 - EXERCÍCIOS RESOLVIDOS, 41
5.6 - EXERCÍCIOS PROPOSTOS, 44

CAPÍTULO 6
NOÇÕES SOBRE CORRELAÇÃO, 45

6.1 - DIAGRAMA DE DISPERSÃO, 45
6.2 - CORRELAÇÃO POSITIVA E CORRELAÇÃO NEGATIVA, 46
6.3 - COEFICIENTE DE CORRELAÇÃO, 48
6.4 - EXERCÍCIOS RESOLVIDOS, 51
6.5 - EXERCÍCIOS PROPOSTOS, 53

CAPÍTULO 7
NOÇÕES SOBRE REGRESSÃO, 57

7.1 - GRÁFICO DE LINHAS, 57
7.2 - RETA DE REGRESSÃO, 58
7.3 - ESCOLHA DA VARIÁVEL EXPLANATÓRIA, 61
7.4 - TRANSFORMAÇÃO DE VARIÁVEIS, 63
7.5 - EXERCÍCIOS RESOLVIDOS, 66
7.6 - EXERCÍCIOS PROPOSTOS, 70

CAPÍTULO 8
NOÇÕES SOBRE PROBABILIDADE, 71

8.1 - QUESTÕES BÁSICAS, 71
8.2 - PROBABILIDADE CONDICIONAL, 72
8.3 - EVENTOS INDEPENDENTES, 73
8.4 - TEOREMA DO PRODUTO, 74
8.5 - TEOREMA DA SOMA, 75
8.6 - EXERCÍCIOS RESOLVIDOS, 76
8.7 - EXERCÍCIOS PROPOSTOS, 77

CAPÍTULO 9
DISTRIBUIÇÃO BINOMIAL, 79

9.1 - VARIÁVEL ALEATÓRIA, 79
9.2 - DISTRIBUIÇÃO DISCRETA, 80
9.3 - DISTRIBUIÇÃO BINOMIAL, 81
9.4 - MÉDIA E VARIÂNCIA NA DISTRIBUIÇÃO BINOMIAL, 84
9.5 - EXERCÍCIOS RESOLVIDOS, 84
9.6 - EXERCÍCIOS PROPOSTOS, 87

CAPÍTULO 10
DISTRIBUIÇÃO NORMAL, 89

10.1 - CARACTERÍSTICAS GERAIS, 90
10.2 - DISTRIBUIÇÃO NORMAL REDUZIDA, 90
10.3 - PROBABILIDADES NA DISTRIBUIÇÃO NORMAL, 93
10.4 - APROXIMAÇÃO NORMAL DA BINOMIAL, 95
10.5 - EXERCÍCIOS RESOLVIDOS, 99
10.6 - EXERCÍCIOS PROPOSTOS, 101

CAPÍTULO 11
TESTE DE χ^2, 103

11.1 - CONCEITOS BÁSICOS, 103
11.2 - PROCEDIMENTOS USUAIS, 105
11.3 - TESTE DE χ^2 PARA ADERÊNCIA, 106
11.4 - TESTE DE χ^2 PARA INDEPENDÊNCIA, 109
11.5 - RESTRIÇÕES AO USO DO TESTE DE χ^2, 112
11.6 - RISCO RELATIVO, 112
11.7 - MEDIDAS DE ASSOCIAÇÃO, 114
11.8 - EXERCÍCIOS RESOLVIDOS, 116
11.9 - EXERCÍCIOS PROPOSTOS, 118

CAPÍTULO 12
TESTE *t*, 121

12.1 - TESTE *t* PARA OBSERVAÇÕES INDEPENDENTES, 121
12.2 - EXEMPLO DE APLICAÇÃO, 122
12.3 - TESTE *t* PARA OBSERVAÇÕES PAREADAS, 124
12.4 - EXEMPLO DE APLICAÇÃO, 125
12.5 - TESTE *t* PARA OBSERVAÇÕES INDEPENDENTES QUANDO AS VARIÂNCIAS SÃO DESIGUAIS, 126
12.6 - EXEMPLO DE APLICAÇÃO, 129
12.7 - TESTE *t* PARA O COEFICIENTE DE CORRELAÇÃO, 130
12.8 - EXERCÍCIOS RESOLVIDOS, 131
12.9 - EXERCÍCIOS PROPOSTOS, 134

CAPÍTULO 13
ANÁLISE DE VARIÂNCIA, 137

13.1 - ANÁLISE DE VARIÂNCIA PARA EXPERIMENTOS AO ACASO, 138
13.2 - TESTE DE TUKEY PARA COMPARAÇÃO DE MÉDIAS, 142
13.3 - ANÁLISE DE VARIÂNCIA COM NÚMERO DIFERENTE DE REPETIÇÕES, 143
13.4 - EXERCÍCIOS RESOLVIDOS, 147
13.5 - EXERCÍCIOS PROPOSTOS, 150

CAPÍTULO 14
INTERVALO DE CONFIANÇA, 153

14.1 - ERRO PADRÃO DA MÉDIA, 153
14.2 - INTERVALO DE CONFIANÇA, 155
14.3 - ALGUNS PONTOS BÁSICOS, 156
14.4 - EXERCÍCIOS RESOLVIDOS, 157
14.5 - EXERCÍCIOS PROPOSTOS, 158

CAPÍTULO 15
ELEMENTOS DE MATEMÁTICA, 159

15.1 - SOMATÓRIOS, 159
15.2 - ANÁLISE COMBINATÓRIA, 161
15.3 - EQUAÇÃO DA RETA, 162
15.4 - LOGARITMOS, 163

EXERCÍCIOS DE REVISÃO, 165

APÊNDICES

Tabelas, 169
Respostas aos Exercícios Propostos, 183
Sugestões para Leitura, 189
Referências Bibliográficas, 191
Índice Remissivo, 193

CAPÍTULO
1

Noções Básicas

Nas áreas médica e biológica coletam-se dados de pessoas, de animais experimentais e de fenômenos físicos e químicos. Interessam aos pesquisadores dessas áreas dados sobre mortalidade infantil, eficiência de medicamentos, incidência de doenças, causas de morte etc. Os dados referem-se a *variáveis*, que são classificadas, em Estatística, como qualitativas, ordinais e quantitativas.

1.1 - VARIÁVEIS

Uma variável é *qualitativa* quando os dados podem ser distribuídos em categorias mutuamente exclusivas. Assim, sexo é uma variável qualitativa porque permite distinguir duas categorias, masculino e feminino. Também são qualitativas as variáveis cor, causa de morte, grupo sangüíneo etc.

Uma variável é *ordinal* quando os dados podem ser distribuídos em categorias mutuamente exclusivas que têm ordenação natural. Assim, grau de instrução é uma variável ordinal porque as pessoas podem ser distribuídas em categorias mutuamente exclusivas, na seguinte ordem: primário, secundário e superior. Também são ordinais as variáveis aparência, *status* social, estágio da doença etc.

Uma variável é *quantitativa* quando é expressa por números. São variáveis quantitativas: idade, estatura, peso corporal etc.

1.2 - Apuração de Dados

Os dados são registrados em fichas, com várias outras informações. Para obter apenas os dados é preciso fazer uma *apuração*. Se a variável é qualitativa ou ordinal, a apuração resume-se a simples contagem. Por exemplo, para obter o número de nascidos vivos de cada sexo, é preciso tomar os prontuários e escrever, numa folha de papel:

Masculino
Feminino

Depois, é preciso examinar os prontuários e fazer um traço, na linha correspondente a um dos sexos, toda vez que o prontuário registrar que o nascido vivo é desse sexo. No exemplo, cada traço representa um nascido vivo e cada quadrado, cortado pela diagonal, representa cinco nascidos vivos. O total é dado pelo número de traços em cada linha.

Se a variável é quantitativa, a apuração consiste em anotar cada valor observado. Por exemplo, para apurar dados de peso ao nascer, basta escrever os pesos numa folha de papel. O número do prontuário, escrito ao lado do peso ao nascer, facilita a posterior verificação da apuração. Veja o exemplo.

Nº do prontuário	Peso ao nascer
10 525	3,25
10 526	2,00
.	.
.	.
.	.
10 624	3,20

1.3 - População e Amostra

Entende-se por *população* o conjunto de elementos que têm, em comum, determinada característica. Todo subconjunto não vazio e com menor número de elementos do que a população constitui uma *amostra* dessa população. As populações podem ser *finitas*, como o conjunto de alunos de uma escola em determinado ano, ou *infinitas*, como o número de vezes que se pode jogar um dado.

Para certas finalidades, as populações finitas muito grandes são consideradas infinitas. Como exemplo, considere as pessoas do sexo masculino, com mais de 35 anos de idade, residentes na cidade de São Paulo. O número dessas pessoas é matematicamente finito, mas tão grande que um pesquisador, ao analisar uma amostra de 500 pessoas, pode considerar a população como infinita.

Quando são coletadas informações de toda a população, diz-se que foi feito um *recenseamento*. Censo é o conjunto de dados obtidos através de recenseamento. Quando são coletadas informações de apenas parte da população, diz-se que foi feita uma *amostragem*. Amostra é tanto a parte retirada da população para estudo como, também, o conjunto de dados obtidos nessa parte da população.

Os pesquisadores trabalham com amostras, por vários motivos. Primeiro, é fato que as populações infinitas só podem ser estudadas através de amostras. Por exemplo, por maior que seja o número de vezes que uma pessoa possa pesar um corpo sólido, o resultado será sempre uma amostra porque, teoricamente, todo corpo pode ser pesado um número infinito de vezes.

Depois, as populações finitas muito grandes só podem ser estudadas através de amostras. Por exemplo, o número de cobaias existentes no mundo em determinado período de tempo é, matematicamente, finito, mas as pesquisas que usam cobaias só podem ser feitas com amostras, porque nenhum pesquisador dispõe de todas as cobaias do mundo para seu trabalho.

Finalmente, o estudo cuidadoso de uma amostra tem mais valor científico do que o estudo sumário de toda a população. Por exemplo, para estudar o efeito do flúor sobre a prevenção de cáries em crianças, é melhor submeter uma amostra de crianças a exames periódicos minuciosos, do que examinar rapidamente todas as crianças antes, e determinado tempo após o uso do flúor.

1.4 - Técnicas de Amostragem

Definida a população, é preciso estabelecer a *técnica de amostragem*, isto é, o procedimento que será adotado para escolher os elementos que irão compor a amostra. Conforme a técnica utilizada, tem-se um tipo de amostra.

1.4.1 - *Amostra casual simples*

A amostra casual simples é composta por elementos retirados ao acaso da população. Então todo elemento da população tem igual proba-

bilidade de ser escolhido para a amostra. Um exemplo ajuda a entender essa técnica de amostragem.

Imagine que um professor quer obter uma amostra casual simples dos alunos de sua escola. Para isso, pode organizar um sorteio com fichas numeradas, de zero a nove. Para fazer o sorteio, o professor retira uma ficha de uma urna e anota o número. Esse número será o primeiro dígito do número do aluno que será sorteado para a amostra. Feito isso, o professor recoloca a ficha na urna, mistura, retira outra ficha e anota o número, que será o segundo dígito do número do aluno que será sorteado para a amostra. Esse procedimento deve ser repetido até que sejam retirados todos os dígitos do número do aluno sorteado.

Se a escola tem, por exemplo, 832 alunos, os números dos alunos têm três dígitos. Para sortear um aluno, é preciso retirar três fichas da urna, uma de cada vez, sempre lembrando que a ficha retirada deve ser recolocada na urna, antes de nova retirada. O número de um dos alunos sorteados poderia ser, por exemplo, 377, assim obtido:

Primeira ficha: 3
Segunda ficha: 7
Terceira ficha: 7

É claro que devem ser desprezados números maiores do que 832 (se a escola tem 832 alunos, nenhum aluno recebeu número maior do que 832), números que já foram sorteados e o número 000. O professor sorteia tantos números quantos são os alunos que ele quer na amostra.

1.4.2 - *Amostra sistemática*

Na amostra sistemática os elementos são escolhidos não por acaso, mas por um *sistema*. No exemplo, o professor terá organizado uma amostra sistemática se, em lugar de sortear os alunos, chamar para a amostra todo aluno com número terminado em determinado dígito. Veja o esquema dado em seguida. O professor chamou, para a amostra, todos os alunos com números terminados em zero, assinalados no esquema com asteriscos. Então organizou uma amostra sistemática.

Quando a população está organizada, é mais fácil obter uma amostra sistemática do que uma amostra casual simples. Por exemplo, para obter uma amostra de 2% dos prontuários dos pacientes de uma clínica, é mais fácil pegar o último de cada 50 prontuários do que fazer um sorteio até conseguir 2% do total de prontuários.

As amostras sistemáticas são muito usadas, mas exigem especial preocupação com o sistema de seleção. Por exemplo, se os elementos da população estão em fila, não se deve selecionar os "primeiros", ou os "últimos", nem mesmo "os do meio"; é preciso percorrer toda a fila e escolher, por exemplo, o décimo de cada grupo de dez.

Nº	Nome	Nº	Nome	Nº	Nome
1		21		41	
.		.		.	
.		.		.	
.		.		.	
*10		*30		*50	
.		.		.	
.		.		.	
.		.		.	
*20		*40		*60	

1.4.3 - *Amostra estratificada*

A amostra estratificada é composta por elementos provenientes de todos os *estratos* da população. No exemplo, se o professor considera que alunos de diferentes séries apresentam reais diferenças, cada série é um estrato. O professor deve, então, obter uma amostra de cada série (estrato) e depois reunir todas as amostras em uma só. Esta amostra final é estratificada.

Devem ser obtidas amostras estratificadas sempre que a população for constituída por diferentes estratos. Por exemplo, se as pessoas que moram nos vários bairros de uma cidade são diferentes, cada bairro é um estrato. Para obter uma amostra de pessoas dessa cidade, seria razoável obter uma amostra de cada bairro e depois reunir as informações numa amostra estratificada.

1.4.4 - *Amostra de conveniência*

A amostra de conveniência é formada por elementos que o pesquisador reuniu simplesmente porque dispunha deles. Então, se o professor tomar os alunos de sua classe como amostra de toda a escola, estará usando uma amostra de conveniência.

Os estatísticos têm muitas restrições ao uso de amostras de conveniência. Mesmo assim, as amostras de conveniência são comuns na área de saúde, onde se fazem pesquisas com pacientes de uma só clínica ou de um só hospital. Mais ainda, as amostras de conveniência constituem, muitas vezes, a única maneira de estudar determinado problema.

De qualquer forma, o pesquisador que utiliza amostras de conveniência precisa de muito senso crítico. Os dados podem ser tendenciosos. Por exemplo, para estimar a probabilidade de morte por desidratação não se deve recorrer aos dados de um hospital. Como só são internados os casos graves, é possível que a mortalidade entre pacientes internados seja muito maior do que entre pacientes não-internados. Conseqüentemente,

a amostra de conveniência — constituída, neste exemplo, por pacientes internados no hospital — seria tendenciosa.

Finalmente, o pesquisador que trabalha com amostras sempre pretende fazer *inferência*, isto é, estender os resultados da amostra para toda a população. Então é muito importante caracterizar bem a amostra e estender os resultados obtidos na amostra apenas para a população de onde a amostra proveio.

1.5 - Exercícios Resolvidos

1.5.1 - *Os prontuários dos pacientes de um hospital estão organizados em um arquivo, por ordem alfabética. Qual é a maneira mais rápida de amostrar 1/3 do total de prontuários?*

Seleciona-se, para a amostra, um de cada três prontuários ordenados (por exemplo, o terceiro de cada três).

1.5.2 - *Um pesquisador tem dez gaiolas que contêm, cada uma, seis ratos. Como o pesquisador pode selecionar dez ratos para uma amostra?*

O pesquisador pode usar a técnica de amostragem estratificada, isto é, sortear um rato de cada gaiola para compor a amostra.

1.5.3 - *Para levantar dados sobre o número de filhos por casal, em uma comunidade, um pesquisador organizou um questionário que enviou, pelo correio, a todas as residências. A resposta ao questionário era facultativa, pois o pesquisador não tinha condições de exigir a resposta. Nesse questionário perguntava-se o número de filhos por casal morador na residência. Você acha que os dados assim obtidos têm algum tipo de tendenciosidade?*

Neste caso, é razoável esperar os seguintes tipos de tendenciosidade: a) os casais com muitos filhos responderiam, pensando na possibilidade de algum tipo de ajuda, como instalação de uma creche no bairro; b) os casais que recentemente tiveram o primeiro filho também responderiam; c) muitos dos casais que não têm filhos não responderiam.

1.5.4 - *Um pesquisador pretende levantar dados sobre o número de moradores por domicílio, usando a técnica de amostragem sistemática. Para isso, o pesquisador visitará cada domicílio selecionado. Se nenhuma pessoa estiver presente na ocasião da visita, o pesquisador excluirá o domicílio da amostra. Esta última determinação introduz tendenciosidade. Por quê?*

Nos domicílios onde moram muitas pessoas, será mais fácil o pesquisador encontrar pelo menos uma pessoa, por ocasião de sua visita. Então é

razoável admitir que os domicílios com poucos moradores têm maior probabilidade de serem excluídos da amostra.

1.5.5 - *Muitas pessoas acreditam que as famílias se tornaram menores. Suponha que, para estudar essa questão, foi selecionada uma amostra de 2.000 casais e perguntou-se quantos filhos eles tinham, quantos filhos tinham seus pais e quantos filhos tinham seus avós. O procedimento introduz tendenciosidade nos dados. Por quê?*

Os casais de gerações anteriores que não tiveram filhos não têm possibilidade de ser selecionados para a amostra. Por outro lado, os casais de gerações anteriores que tiveram muitos filhos terão grande probabilidade de ser amostrados.

1.6 - Exercícios Propostos

1.6.1 - *Dada uma população de 4 pessoas, Antônio, Luís, Pedro e Carlos, quantas amostras casuais simples de tamanho 2 podem ser obtidas? Quais são essas amostras?*

1.6.2 - *Dada uma população de 8 elementos, A, B, C, D, E, F, G e H, descreva três formas diferentes de obter uma amostra sistemática de 4 elementos.*

1.6.3 - *Dada uma população de 40 alunos, descreva uma forma de obter uma amostra casual simples de 6 alunos.*

1.6.4 - *Organize uma lista com 10 nomes de pessoas em ordem alfabética. Depois descreva uma forma de obter uma amostra sistemática de 5 nomes.*

1.6.5 - *Em uma pesquisa de mercado para serviços odontológicos tomou-se a lista telefônica, onde os nomes dos assinantes estão organizados em ordem alfabética do último sobrenome, e se amostrou o décimo de cada 10 assinantes. Critique esse procedimento.*

CAPÍTULO
2

Apresentação de Dados em Tabelas

Os dados devem ser apresentados em tabelas construídas de acordo com as normas técnicas ditadas pela Fundação Instituto Brasileiro de Geografia e Estatística (Fundação IBGE).

2.1 - Componentes das Tabelas

As tabelas têm título, corpo, cabeçalho e coluna indicadora. O *título* explica o que a tabela contém. O *corpo* é formado pelas linhas e colunas de dados. O *cabeçalho* especifica o conteúdo das colunas, e a *coluna indicadora* especifica o conteúdo das linhas. Como exemplo, veja a Tabela 2.1.

Tabela 2.1
Casos registrados de intoxicação humana, segundo a causa determinante. Brasil, 1993

Causa	Freqüência
Acidente	29.601
Abuso	2.604
Suicídio	7.965
Profissional	3.735
Outras	1.959
Ignorada	1.103

Fonte: MS/FIOCRUZ/SINITOX

Na Tabela 2.1, observe o título:

Casos registrados de intoxicação humana, segundo a causa determinante. Brasil, 1993

O cabeçalho é constituído pelas palavras:

Causa	Freqüência

A coluna indicadora é constituída pelas especificações:

Acidente
Abuso
Suicídio
Profissional
Outras
Ignorada

O corpo da tabela é formado pelos números:

29.601
2.604
7.965
3.735
1.959
1.103

Toda tabela deve ser delimitada por traços horizontais. Podem ser feitos traços verticais para *separar* as colunas, mas não devem ser feitos traços verticais para *delimitar* a tabela. O cabeçalho é separado do corpo por um traço horizontal.

As tabelas podem apresentar, além das freqüências, as freqüências relativas e o total. Para obter a *freqüência relativa* de uma dada categoria, divide-se a freqüência dessa categoria pela soma das freqüências. O resultado, multiplicado por 100, é uma porcentagem. O *total* da coluna é escrito entre dois traços horizontais. Veja a Tabela 2.2.

As tabelas podem conter fonte, notas e chamadas. A *fonte* dá indicação da entidade, ou do pesquisador, ou dos pesquisadores que publicaram ou forneceram os dados. Como exemplo, veja a Tabela 2.1. A fonte é MS/FIOCRUZ/SINITOX, que publicou os dados.

As *notas* devem esclarecer aspectos relevantes do levantamento dos dados ou da apuração. Observe a Tabela 2.3. A nota informa que só foram apurados nascimentos ocorridos no ano de registro.

Tabela 2.2

Casos registrados de intoxicação humana, segundo a causa determinante. Brasil, 1993

Causa	Freqüência	Freqüência relativa
Acidente	29.601	63,03
Abuso	2.604	5,54
Suicídio	7.965	16,96
Profissional	3.735	7,95
Outros	1.959	4,17
Ignorada	1.103	2,35
Total	46.967	100,00

Tabela 2.3

Nascidos vivos registrados segundo o ano do registro

Ano do registro	Freqüência
1984	2 559 038
1985	2 619 604
1986	2 779 253

Fonte: IBGE (1988)
Nota: Nascimentos ocorridos no ano de registro

As *chamadas* dão esclarecimentos sobre os dados. Devem ser feitas através de algarismos arábicos escritos entre parênteses, e colocadas à direita da coluna.

2.2 - Tabelas de Contingência

Muitas vezes os elementos da amostra ou da população são classificados de acordo com dois fatores. Os dados devem então ser apresentados em *tabelas de contingência*, isto é, em tabelas de dupla entrada, cada entrada relativa a um dos fatores. Como exemplo, veja a Tabela 2.4, que apresenta o número de nascidos vivos registrados. Note que eles estão classificados segundo dois fatores: o ano de registro e o sexo.

As tabelas de contingência podem apresentar freqüências relativas, além de freqüências. As freqüências relativas dão estimativas de *riscos*, isto é, dão estimativas das probabilidades de dano. Veja a Tabela 2.5.

Tabela 2.4

Nascidos vivos registrados segundo o ano de registro e o sexo

Ano de registro	Sexo		Total
	Masculino	Feminino	
1984	1 307 758	1 251 280	2 559 038
1985	1 339 059	1 280 545	2 619 604
1986	1 418 050	1 361 203	2 779 253

Fonte: IBGE (1988)
Nota: Nascimentos ocorridos no ano de registro

Tabela 2.5

Recém-nascidos segundo a época do ataque de rubéola na gestante e a condição de normal ou defeituoso

Época do ataque	Condição		Total	Freqüência relativa de defeituosos
	Normal	Defeituoso		
Até o 3° mês.....	36	14	50	28,0%
Depois do 3° mês	51	3	54	5,6%

Fonte: HILL et alii (1958)

As freqüências relativas apresentadas na Tabela 2.5 estimam o risco de um recém-nascido ser defeituoso em função da época em que a gestante foi atacada de rubéola. Note que a freqüência relativa de defeituosos (risco) é maior quando a gestante foi atacada de rubéola no primeiro trimestre da gestação. Diz-se então que a época do ataque de rubéola é um *fator de risco* na ocorrência de recém-nascidos defeituosos.

2.3 - Tabelas de Distribuição de Freqüências

As tabelas com grande número de dados são cansativas e não dão ao leitor visão rápida e global do fenômeno. Para isso, é preciso que os dados estejam organizados em uma *tabela de distribuição de freqüências*. Nesta seção se explica, passo a passo, a construção desse tipo de tabela usando, como exemplo, os dados da Tabela 2.6.

Imagine que, para dar uma idéia geral sobre peso ao nascer de nascidos vivos, o pesquisador irá apresentar não os pesos observados, mas o número de nascidos vivos por faixas de peso. Deve, então, construir uma tabela de distribuição de freqüências.

Tabela 2.6

Peso ao nascer de nascidos vivos, em quilogramas

2,522	3,200	1,900	4,100	4,600	3,400
2,720	3,720	3,600	2,400	1,720	3,400
3,125	2,800	3,200	2,700	2,750	1,570
2,250	2,900	3,300	2,450	4,200	3,800
3,220	2,950	2,900	3,400	2,100	2,700
3,000	2,480	2,500	2,400	4,450	2,900
3,725	3,800	3,600	3,120	2,900	3,700
2,890	2,500	2,500	3,400	2,920	2,120
3,110	3,550	2,300	3,200	2,720	3,150
3,520	3,000	2,950	2,700	2,900	2,400
3,100	4,100	3,000	3,150	2,000	3,450
3,200	3,200	3,750	2,800	2,720	3,120
2,780	3,450	3,150	2,700	2,480	2,120
3,155	3,100	3,200	3,300	3,900	2,450
2,150	3,150	2,500	3,200	2,500	2,700
3,300	2,800	2,900	3,200	2,480	-
3,250	2,900	3,200	2,800	2,450	-

Primeiro, é preciso definir as faixas de peso que recebem, tecnicamente, o nome de *classes*. Observe os dados apresentados na Tabela 2.6. O menor valor é 1,570kg e o maior valor é 4,600kg. Podem então ser definidas classes de 1,5 a 2,0kg, de 2,0 a 2,5kg, e assim por diante, como mostra o esquema dado a seguir:

1,5 ⊢ 2,0
2,0 ⊢ 2,5
2,5 ⊢ 3,0
3,0 ⊢ 3,5
3,5 ⊢ 4,0
4,0 ⊢ 4,5
4,5 ⊢ 5,0

Na classe de 1,5 a menos de 2,0kg são colocados desde nascidos com 1,5kg até os que nasceram com 1,999kg; na classe de 2,0 a menos de 2,5kg são colocados desde nascidos com 2,0kg até os que nasceram com 2,499kg, e assim por diante. Logo, cada classe cobre um intervalo de 0,5kg, ou seja, cada *intervalo de classe* é de 0,5kg. É mais fácil trabalhar com intervalos de classe iguais. A distribuição das freqüências, obtida a partir da Tabela 2.6, é dada a seguir.

Classe	Freqüência	
1,5 ⊢ 2,0	⊏	= 3
2,0 ⊢ 2,5	⊠⊠⊠|	= 16
2,5 ⊢ 3,0	⊠⊠⊠⊠⊠⊠|	= 31
3,0 ⊢ 3,5	⊠⊠⊠⊠⊠⊠⊏	= 34
3,5 ⊢ 4,0	⊠⊠|	= 11
4,0 ⊢ 4,5	⊏	= 4
4,5 ⊢ 5,0	|	= 1

Denominam-se *extremos de classe* os limites dos intervalos de classe. Deve ficar muito claro se os valores iguais aos extremos devem ou não ser incluídos na classe. Recomenda-se adotar a notação 1,5 ⊢ 2,0, 2,0 ⊢ 2,5 etc. Isto significa que o intervalo é *fechado* à *esquerda*, isto é, pertencem à classe os valores iguais ao extremo inferior (por exemplo, 1,5 na primeira classe). Também significa que o intervalo é *aberto* à *direita*, isto é, não pertencem à classe os valores iguais ao extremo superior (por exemplo, 2,0 na primeira classe).

Numa tabela de distribuição de freqüências também podem ser apresentados os pontos médios de classe. O *ponto médio* é dado pela soma dos extremos da classe, dividida por 2. Para a classe 1,5 ⊢ 2,0, o ponto médio é:

$$\frac{1,5 + 2,0}{2} = 1,75$$

Uma tabela típica de distribuição de freqüências tem, então, três colunas: a da esquerda, onde estão escritas as classes; a do meio, onde estão escritos os pontos médios; e a da direita, onde estão escritas as freqüências, isto é, o número de elementos de cada classe. Veja a Tabela 2.7.

Tabela 2.7

Nascidos vivos segundo o peso ao nascer, em quilogramas

Classe	Ponto médio	Freqüência
1,5 ⊢ 2,0	1,75	3
2,0 ⊢ 2,5	2,25	16
2,5 ⊢ 3,0	2,75	31
3,0 ⊢ 3,5	3,25	34
3,5 ⊢ 4,0	3,75	11
4,0 ⊢ 4,5	4,25	4
4,5 ⊢ 5,0	4,75	1

Nem sempre estão definidos o extremo inferior da primeira classe ou o extremo superior da última classe. Observe a Tabela 2.8. O extremo superior da última classe não está definido. Esta tabela também exemplifica o uso de intervalos de classe diferentes.

Tabela 2.8

Mulheres com 30 anos de idade segundo a pressão sangüínea sistólica, em milímetros de mercúrio

Classe	Ponto médio	Freqüência
90 ⊢ 100	95	6
100 ⊢ 105	102,5	11
105 ⊢ 110	107,5	12
110 ⊢ 115	112,5	17
115 ⊢ 120	117,5	18
120 ⊢ 125	122,5	11
125 ⊢ 130	127,5	9
130 ⊢ 135	132,5	6
135 ⊢ 140	137,5	4
140 ⊢ 150	145	4
150 ⊢ 160	155	1
160 e mais	...	1

As tabelas de distribuição de freqüências mostram a distribuição da variável, mas perdem em exatidão. Isto porque todos os dados passam a ser representados pelo ponto médio da classe a que pertencem. Por exemplo, a Tabela 2.8 mostra que seis mulheres apresentaram pressão sangüínea sistólica com o ponto médio igual a 95, mas não dá informação exata sobre a pressão de cada uma delas.

O número de classes deve ser escolhido pelo pesquisador, em função do que ele quer mostrar. Em geral, convém estabelecer de 5 a 20 classes. Se o número de classes for demasiado pequeno (por exemplo, 3), perde-se muita informação. Se o número de classes for grande (por exemplo, 30), têm-se pormenores desnecessários. Mas não existe um número "ideal" de classes, embora existam até fórmulas para estabelecer quantas classes devem ser construídas. Uma dessas fórmulas é a seguinte:

$$k = 1 + 3{,}222 \cdot \log n,$$

onde n é o número de dados. O número de classes é um inteiro próximo de k.

Para entender como se aplica esta fórmula, veja a Tabela 2.6. Como $n = 100$, tem-se que

$$k = 1 + 3{,}222 \cdot \log 100 = 7{,}444,$$

ou seja, deveriam ter sido construídas 7 ou 8 classes.

É importante deixar claro, aqui, que o resultado obtido por esta fórmula pode ser usado como referência, mas cabe ao pesquisador determinar o número de classes que pretende organizar. Finalmente, quando se constrói uma tabela de distribuição de freqüências, é melhor usar, como extremos de classes, números fáceis de trabalhar. No caso do peso ao nascer dos nascidos vivos, foram definidas 7 classes e foram estabelecidos extremos com valores fáceis, como 1,5 e 2,0.

2.4 - Exercícios Resolvidos

2.4.1 - ***De acordo com o IBGE (1988), a distribuição dos suicídios ocorridos no Brasil em 1986, segundo a causa atribuída, foi a seguinte: 263 por alcoolismo, 198 por dificuldade financeira, 700 por doença mental, 189 por outro tipo de doença, 416 por desilusão amorosa e 217 por outras causas. Apresente essa distribuição em uma tabela.***

A resposta para este exercício é a Tabela 2.9.

Tabela 2.9

Suicídios ocorridos no Brasil em 1986, segundo a causa atribuída

Causa atribuída	Freqüência	Percentagem
Alcoolismo	263	13,26
Dificuldade financeira	198	9,98
Doença mental	700	35,30
Outro tipo de doença	189	9,53
Desilusão amorosa	416	20,98
Outras	217	10,94

Fonte: IBGE (1988)

2.4.2 - ***Construa uma tabela de distribuição de freqüências para apresentar os dados da Tabela 2.10.***

Para determinar o número de classes pode ser usada a fórmula:

$$k = 1 + 3{,}222 \cdot \log n,$$

onde n é 49. Então,

$$k = 1 + 3{,}222 \cdot 1{,}6902 = 6{,}4458.$$

Tabela 2.10

Pressão arterial, em milímetros de mercúrio, de cães adultos anestesiados e após laparotomia

130,0	105,0	120,0	111,5	99,0	116,0	82,5
107,5	125,0	100,0	107,5	120,0	143,0	115,0
135,0	130,0	135,0	127,5	90,5	104,5	136,5
100,0	145,0	125,0	104,5	101,5	102,5	101,5
134,5	158,5	110,0	102,5	90,5	107,5	124,0
121,5	135,0	102,0	119,5	115,5	125,5	117,5
107,5	140,0	121,5	107,5	113,0	93,0	103,5

Fonte: ARAÚJO e HOSSNE (1977)

De acordo com a fórmula, podem ser constituídas 6 ou 7 classes. Como o menor valor observado é 82,5 e o maior valor é 158,5, é razoável construir classes com intervalos iguais a 10, a partir de 80. O número de classes será, então, 8, um pouco maior do que o estabelecido pela fórmula.

Tabela 2.11

Cães adultos anestesiados e após laparotomia segundo a pressão arterial, em milímetros de mercúrio

Classe	Ponto médio	Freqüência
80 ⊢ 90	85	1
90 ⊢ 100	95	4
100 ⊢ 110	105	16
110 ⊢ 120	115	8
120 ⊢ 130	125	9
130 ⊢ 140	135	7
140 ⊢ 150	145	3
150 ⊢ 160	155	1

2.5 - Exercícios Propostos

2.5.1 - *De acordo com o IBGE (1988), em 1986 ocorreram, em acidentes de trânsito, 27 306 casos de vítimas fatais, assim distribuídos: 11 712 pedestres, 7 116 passageiros e 8 478 condutores. Faça uma tabela para apresentar esses dados. Apresente também as freqüências relativas e o total.*

2.5.2 - *De posse da Tabela 2.12, calcule as freqüências relativas de não-sobreviventes.*

Tabela 2.12

Pacientes com câncer de mama segundo a faixa de idade por ocasião do diagnóstico e a sobrevivência após três anos

Faixa de idade	Sobrevivência	
	Sim	Não
Menos de 50 anos	11	6
De 50 a 70 anos	18	8
Mais de 70 anos	15	9

Fonte: MORRISON (1973)

2.5.3 - ***De posse da Tabela 2.13, calcule as freqüências relativas em cada linha, isto é, calcule a proporção de estabelecimentos de saúde, públicos e particulares, de cada espécie.***

Tabela 2.13

Estabelecimentos de saúde, públicos e particulares, por espécie. Brasil, 1985

Espécie	Estabelecimentos	
	Públicos	Particulares
Hospital	1 002	5 132
Pronto-socorro	150	156
Policlínicas	1 531	6 136
Outros (1)	14 393	472

Fonte: IBGE (1988)

(1) Inclui postos de saúde, centros de saúde e unidades mistas

2.5.4 - ***Construa uma tabela de distribuição de freqüências para apresentar os dados da Tabela 2.14, usando intervalos de classes iguais. Depois faça outra tabela, com os seguintes intervalos: 1 dia, 2 ou 3 dias, de 4 a 7 dias, de 8 a 14 dias, mais de 14 dias.***

Tabela 2.14

Tempo de internação, em dias, de pacientes acidentados no trabalho, em um dado hospital

7	8	1	7	13	6
12	12	3	17	4	2
4	15	2	14	3	5
10	8	9	8	5	3
2	7	14	12	10	8
1	6	4	7	7	11

CAPÍTULO
3

Apresentação de Dados em Gráficos

Existem normas nacionais para a construção de gráficos, ditadas pela Fundação IBGE. Assim, todo gráfico deve apresentar título e escala. O *título* pode ser colocado tanto acima como abaixo do gráfico. As *escalas* devem crescer da esquerda para a direita, e de baixo para cima. As *legendas explicativas* devem ser colocadas, de preferência, à direita do gráfico.

3.1 - Gráfico de Barras

O gráfico de barras é usado para apresentar variáveis qualitativas ou ordinais. Para fazer um *gráfico de barras*, primeiro se traça o sistema de eixos cartesianos. Depois colocam-se, no eixo das abscissas (ou das ordenadas) as categorias da variável em estudo. Em seguida, constroem-se barras retangulares com base no eixo das abscissas (ou das ordenadas) e altura (ou comprimento) igual à freqüência, ou à freqüência.

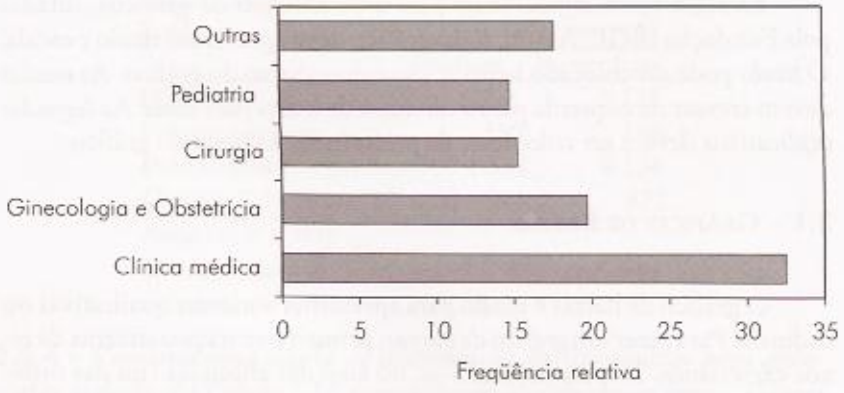
Os dados da Tabela 3.1 estão apresentados em gráfico de barras na Figura 3.1.

Tabela 3.1
Internações em estabelecimentos de saúde, por espécie de clínica — 1992

Espécie de clínica	Freqüência	Freqüência relativa(%)
Médica	6 457 923	32,51
Ginecologia e Obstetrícia	3 918 308	19,73
Cirurgia	3 031 075	15,26
Pediatria	2 943 939	14,82
Outras	3 513 186	17,69

Fonte: IBGE, Diretoria de Pesquisas, Pesquisa de Assistência Médico-Sanitária

Figura 3.1 *Internações em estabelecimentos de saúde, por espécie de clínica. IBGE 1992*



3.2 - Gráfico de Setores

O gráfico de setores também é usado para apresentar variáveis qualitativas ou ordinais. Para fazer um *gráfico de setores*, primeiro se traça uma circunferência que, como se sabe, tem 360°. Essa circunferência representa o total, ou seja, 100%. Dentro dessa circunferência devem ser representadas as categorias da variável em estudo. Para isso, toma-se a freqüência relativa de cada categoria e calcula-se o ângulo central, da seguinte maneira: se 100% correspondem a 360°, uma categoria com freqüência relativa de f% terá um ângulo central x, tal que:

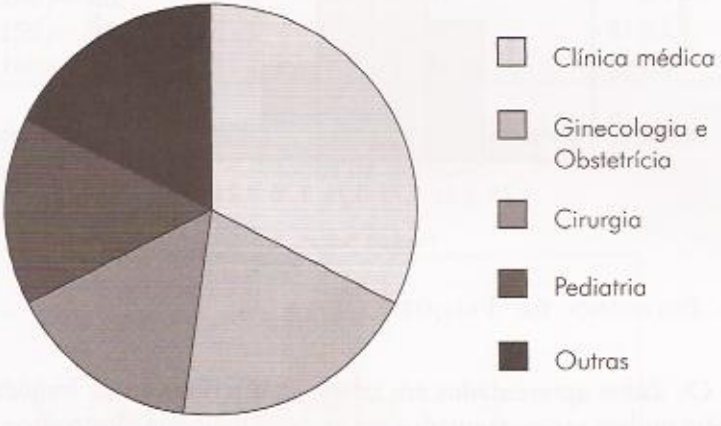
$$100 \rightarrow 360$$
$$f \rightarrow x$$

Logo, o valor do ângulo central x será:

$$x = \frac{360}{100} f$$

O ângulos centrais das demais categorias são obtidos da mesma maneira. Para fazer o gráfico de setores marcam-se, na circunferência, os ângulos calculados, separando-os com o traçado dos raios. Observe o gráfico de setores apresentado na Figura 3.2, feito com os dados da Tabela 3.1.

Figura 3.2 *Internações em estabelecimentos de saúde, por espécie de clínica. IBGE 1992*



3.3 - HISTOGRAMA

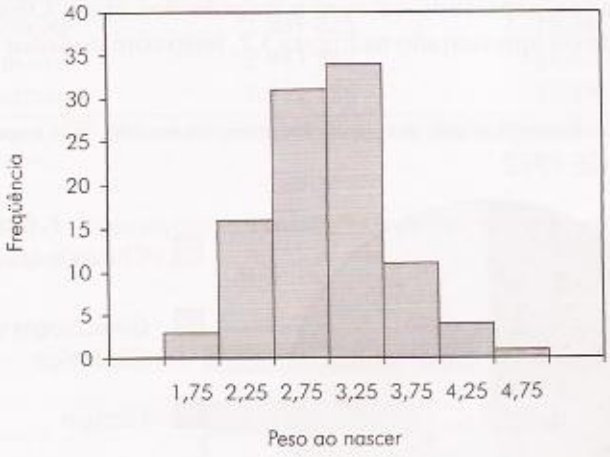
Os dados apresentados em tabelas de distribuição de freqüências são apresentados graficamente em *histogramas*. Para construir um histograma, primeiro se traça o sistema de eixos cartesianos. Depois, se os intervalos de classe são iguais, traçam-se barras retangulares com bases iguais, correspondendo aos intervalos de classe, e com alturas determinadas pelas respectivas freqüências. A Figura 3.3 mostra o histograma feito com a distribuição de freqüências apresentada na Tabela 2.7 do Capítulo 2.

Quando os intervalos de classe são diferentes, para construir um histograma é preciso calcular as densidades de freqüência relativa. Entende-se por *densidade de freqüência relativa* o quociente entre a freqüência relativa e o intervalo de classe, isto é:

$$\text{densidade} = \frac{\text{freqüência relativa}}{\text{intervalo de classe}}$$

Para construir o histograma, desenham-se barras retangulares. As bases são iguais aos intervalos de classe, e as alturas são determinadas pelas respectivas densidades. Veja, como exemplo, a Figura 3.4, feita para apresentar os dados da Tabela 3.2. Note que, para a classe "160 e mais" considerou-se como extremo superior o valor 170. As densidades de classe estão apresentadas na Tabela 3.2.

Figura 3.3 *Nascidos vivos segundo o peso ao nascer, em quilogramas*



3.4 - Polígono de Freqüências

Os dados apresentados em tabela de distribuição de freqüências também podem ser apresentados em gráficos denominados *polígonos de freqüências*. Para fazer esse tipo de gráfico, primeiro se traça o sistema de eixos cartesianos. Depois, se os intervalos de classe são iguais, marcam-se pontos com abscissas iguais aos pontos médios de classes e ordenadas iguais às respectivas freqüências. Se os intervalos de classe são diferentes, marcam-se pontos com abscissas iguais aos pontos médios de classes e ordenadas iguais às respectivas densidades de freqüência relativa.

Para fechar o polígono, unem-se os extremos da figura com o eixo horizontal, nos pontos de abscissas iguais aos pontos médios de uma classe imediatamente inferior à primeira, e de uma classe imediatamente superior à última. Veja o polígono de freqüências apresentado na Figura 3.5, construído para apresentar os dados da Tabela 2.7.

Tabela 3.2
Mulheres com 30 anos de idade segundo a pressão sangüínea sistólica, em milímetros de mercúrio

Classes	Freqüência	Freqüência relativa	Densidade
90 ⊢ 100	6	6	0,6
100 ⊢ 105	11	11	2,2
105 ⊢ 110	12	12	2,4
110 ⊢ 115	17	17	3,4
115 ⊢ 120	18	18	3,6
120 ⊢ 125	11	11	2,2
125 ⊢ 130	9	9	1,8
130 ⊢ 135	6	6	1,2
135 ⊢ 140	4	4	0,8
140 ⊢ 150	4	4	0,4
150 ⊢ 160	1	1	0,1
160 e mais	1	1	0,1

Figura 3.4 *Mulheres com 30 anos de idade segundo a pressão sangüínea sistólica, em milímetros de mercúrio*

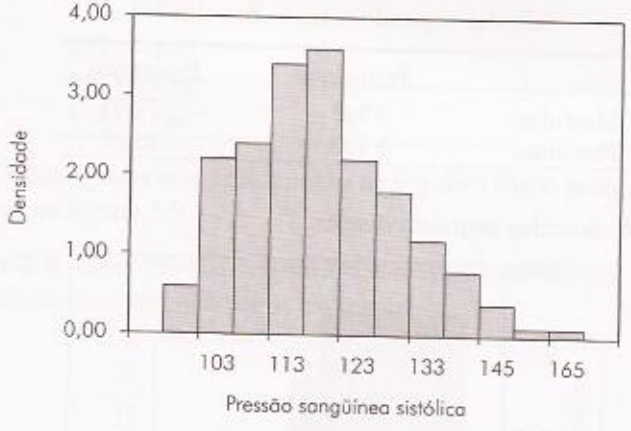
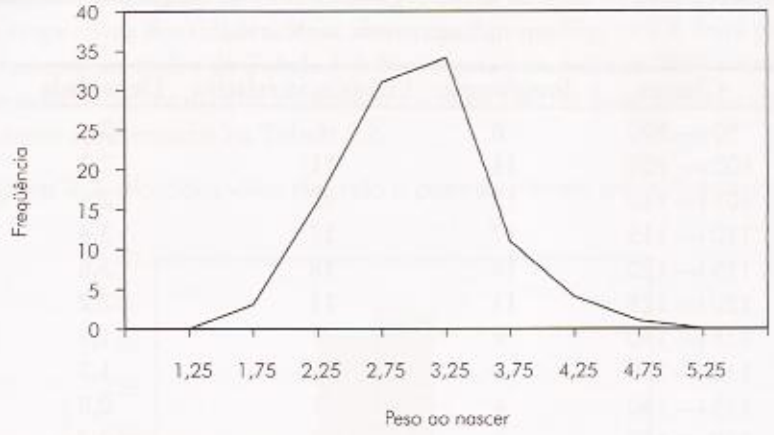


Figura 3.5 *Nascidos vivos segundo o peso ao nascer, em quilogramas*



3.5 - Exercícios Resolvidos

3.5.1 - ***Faça um gráfico de barras e um gráfico de setores para apresentar os dados da Tabela 3.3.***

O gráfico de barras está na Figura 3.6 e o gráfico de setores está na Figura 3.7.

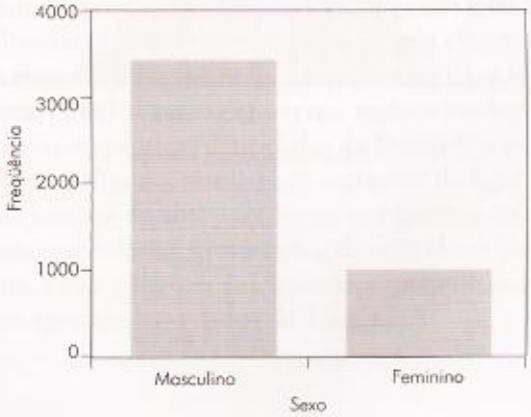
Tabela 3.3

Suicidas segundo o sexo. Brasil, 1986

Sexo	Freqüência	Percentual
Masculino	3 562	74,93
Feminino	1 192	25,07

Fonte: IBGE (1988)

Figura 3.6 *Suicidas segundo o sexo*



3.5.2 - ***Faça um histograma e um polígono de freqüências para apresentar dados da Tabela 3.4.***

Figura 3.7 *Suicidas segundo o sexo*

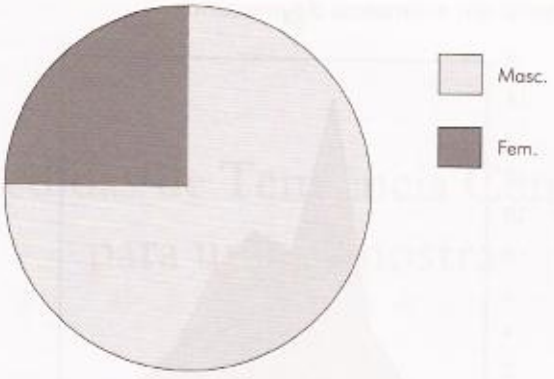


Tabela 3.4

Cães adultos anestesiados e após laparotomia, segundo a pressão arterial, em milímetros de mercúrio

Classe	Ponto médio	Freqüência
80 ⊢ 90	85	1
90 ⊢ 100	95	4
100 ⊢ 110	105	16
110 ⊢ 120	115	8
120 ⊢ 130	125	9
130 ⊢ 140	135	7
140 ⊢ 150	145	3
150 ⊢ 160	155	1

O histograma está apresentado na Figura 3.8, e o polígono de freqüências na Figura 3.9.

Figura 3.8 *Cães adultos anestesiados e após laparotomia, segundo a pressão arterial em milímetros de mercúrio*

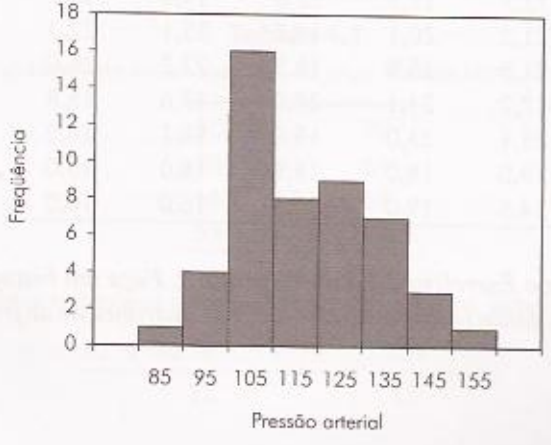
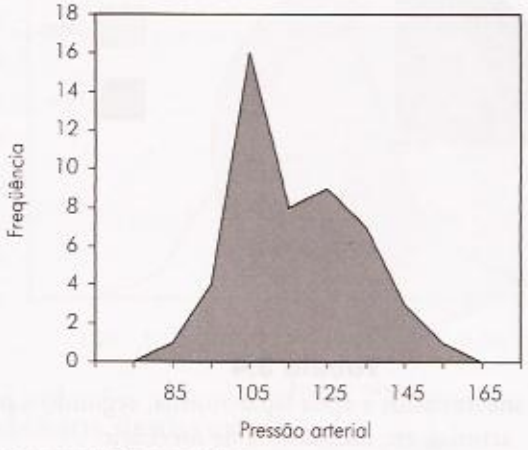


Figura 3.9 *Cães adultos anestesiados e após laparotomia, segundo a pressão arterial em milímetros de mercúrio*



3.6 - Exercícios Propostos

3.6.1 - *Veja o Exercício 2.5.1 do Capítulo 2. Faça um gráfico de barras para apresentar aqueles dados.*

3.6.2 - *Veja o Exercício 2.5.1 do Capítulo 2. Faça um gráfico de setores para apresentar aqueles dados.*

3.6.3 - *Construa uma tabela de distribuição de freqüências com os dados apresentados na Tabela 3.5. Depois faça um polígono de freqüências.*

Tabela 3.5
Peso, em quilogramas, de cães

23,0	19,0	23,8	15,0	20,0
22,7	19,5	22,0	14,9	18,3
21,2	20,1	18,7	15,1	22,3
21,5	25,5	19,5	22,2	24,0
17,0	24,1	28,0	13,6	15,8
28,4	23,0	15,0	16,1	16,0
19,0	18,0	18,8	18,0	15,0
14,5	19,0	20,5	16,0	16,0

3.6.4 - *Veja o Exercício 2.5.4 do Capítulo 2. Faça um histograma para apresentar os dados já dispostos na tabela de distribuição de freqüências.*

CAPÍTULO
4

Medidas de Tendência Central para uma Amostra

Os dados quantitativos, apresentados em tabelas e gráficos, constituem a informação básica do problema em estudo. Mas é conveniente apresentar, além dos dados, medidas que mostrem a informação de maneira resumida. As *medidas de tendência central*, definidas neste Capítulo, dão o valor do ponto em torno do qual os dados se distribuem. São medidas de tendência central: a média aritmética (ou simplesmente média), a mediana e a moda.

4.1 - Média Aritmética

Para obter a *média aritmética* basta somar os valores de todos os dados e dividir o total pelo número deles. Observe a Tabela 4.1.

Tabela 4.1
Peso, em gramas, de ratos machos da raça Wistar com 30 dias de idade

50	62	70
86	60	64
66	77	58
55	82	74

A média aritmética dos dados apresentados na Tabela 4.1 é:

$$\frac{50 + 86 + \ldots + 74}{12} = \frac{804}{12} = 67$$

A média aritmética dá a abscissa do centro de gravidade do conjunto de dados. Para entender esta afirmativa, observe a Figura 4.1, que apresenta os dados da Tabela 4.1. Imagine que o eixo são os braços de uma balança e que cada ponto tem uma unidade de massa. Para haver equilíbrio, é preciso colocar o fulcro da balança no ponto em que se situa a média. Então a média é a abscissa do centro de gravidade.

Figura 4.1 *Distribuição de dados sobre o eixo e a respectiva média*

É conveniente introduzir, neste ponto, a fórmula da média aritmética. A variável em estudo será indicada pela letra maiúscula X, e os valores observados dessa variável serão indicados pela letra minúscula x. Para distinguir um valor do outro, serão usados índices. Então o i-ésimo valor observado de X será indicado por x_i. No exemplo, a variável é peso de ratos, e os valores observados são:

$$x_1 = 50,\ x_2 = 86,\ldots,\ x_{12} = 74$$

A média aritmética, que se representa por $\bar{x}$ (lê-se x-barra ou x-traço), é dada pela soma $x_1 + x_2 + \ldots + x_n$, dividida por n. Escreve-se:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

O símbolo $\sum_{i=1}^{n} x_i$ (lê-se somatório de x_i, i de 1 a n) indica que todos os valores x_i devem ser somados, desde o primeiro (x_1) até o n-ésimo (x_n). Para simplificar, muitas vezes se escreve $\sum x$, mas deve ficar claro quais são os valores de x que devem ser somados. Veja maiores explicações sobre somatórios na seção 15.1 do Capítulo 15.

4.2 - Média de Dados em Tabelas de Distribuição de Freqüências

Se os dados estão em uma tabela de distribuição de freqüências, o cálculo da média é feito de outra forma. Considere os dados apresentados na Tabela 4.2.

Tabela 4.2
Nascidos vivos segundo o peso ao nascer, em quilogramas

Classe	Ponto médio	Freqüência
1,5 ⊢ 2,0	1,75	3
2,0 ⊢ 2,5	2,25	16
2,5 ⊢ 3,0	2,75	31
3,0 ⊢ 3,5	3,25	34
3,5 ⊢ 4,0	3,75	11
4,0 ⊢ 4,5	4,25	4
4,5 ⊢ 5,0	4,75	1

O número de nascidos vivos nessa amostra é:

$$n = 3 + 16 + 31 + 34 + 11 + 4 + 1 = 100$$

Para obter a média dos pesos ao nascer dos nascidos vivos da amostra, multiplica-se o ponto médio de cada classe pela respectiva freqüência, somam-se os produtos e divide-se a soma por n. Então a média é

$$\bar{x} = \frac{1{,}75 \cdot 3 + 2{,}25 \cdot 16 + \ldots + 4{,}25 \cdot 4 + 4{,}75 \cdot 1}{100} = \frac{300{,}00}{100} = 3{,}00$$

Generalizando, considere uma tabela de distribuição de freqüências com k classes. Sejam $x_1, x_2, \ldots, x_k$ os valores dos pontos médios de classe e sejam $f_1, f_2, \ldots, f_k$ as respectivas freqüências, como mostra a Tabela 4.3.

Tabela 4.3
Distribuição de freqüências

Ponto médio	Freqüência
x_1	f_1
x_2	f_2
.	.
.	.
.	.
x_k	f_k

A média dos dados da Tabela 4.3 é dada pela soma $x_1 f_1 + x_2 f_2 + \ldots + x_k f_k$ dividida por n, isto é:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{n}, \qquad \text{onde } n = \sum_{i=1}^{k} f_i$$

4.3 - MEDIANA

Se a amostra é constituída por um número ímpar de dados, a *mediana* é o valor que fica no centro dos dados ordenados. Por exemplo, a mediana dos valores

1, 2, 3, 5 e 9

é 3.

Se a amostra é constituída por um número par de dados, a *mediana* é a média aritmética dos dois valores que ficam na posição central dos dados ordenados. Por exemplo, a mediana dos valores

1, 2, 3, 4, 7 e 9

é a média aritmética dos números 3 e 4, ou seja, a mediana é 3,5.

A mediana dá o valor da abscissa do ponto que delimita metade dos dados. É fácil entender esta afirmativa. Considere os dados apresentados na Tabela 4.1. Para obter a mediana é preciso, primeiro, ordená-los:

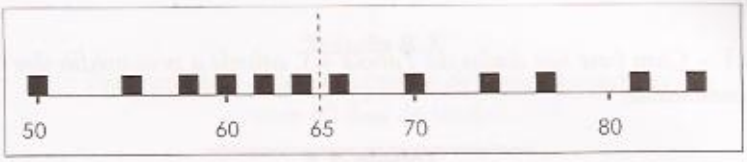
50, 55, 58, 60, 62, 64, 66, 70, 74, 77, 82 e 86

Como nessa amostra o número de dados é par ($n=12$), a mediana é a média aritmética dos dois valores que ocupam a posição central dos dados ordenados, ou seja, a mediana é:

$$\frac{64 + 66}{2} = 65$$

Observe agora o gráfico da Figura 4.2. Nesse gráfico, os dados apresentados na Tabela 4.1 estão representados ao longo de um eixo. A posição da mediana está assinalada pela linha vertical tracejada. É fácil ver que a mediana divide a amostra em dois conjuntos com igual número de dados.

Figura 4.2 *Distribuição de dados sobre o eixo e a respectiva mediana*



4.4 - MODA

A *moda* é o valor que ocorre com maior freqüência. Por exemplo, dados os números

3, 4, 5, 7, 7, 7, 9 e 9,

a moda é 7, porque 7 é o número que ocorre maior número de vezes.

Existem conjuntos de dados que não apresentam moda, porque nenhum valor se repete maior número de vezes, e existem conjuntos de dados com duas ou mais modas. Assim, o conjunto de números

1, 2, 3, 4 e 5

não tem moda, e o conjunto de números

1, 2, 2, 3, 4, 4 e 5

tem duas modas, 2 e 4.

A moda, diferentemente das outras medidas de tendência central, pode ser obtida mesmo que a variável seja qualitativa. Veja os dados apresentados na Tabela 4.8. O sangue tipo O ocorreu com maior freqüência. Então a moda dessa amostra é sangue tipo O.

Tabela 4.4
Indivíduos segundo o tipo de sangue

Tipo de sangue	Freqüência
O	547
A	441
B	123
AB	25

Fonte: GARCIA (1977)

4.5.1 - *Com base nos dados da Tabela 4.5, calcule o peso médio dos ratos em cada idade.*

Tabela 4.5

Peso, em gramas, de ratos machos da raça Wistar segundo a idade, em dias

Número do rato	Idade				
	30	34	38	42	46
1	76,2	95,5	99,2	122,7	134,6
2	81,5	90,0	101,2	125,9	136,2
3	50,0	60,0	62,3	72,2	85,3
4	47,5	50,0	57,5	72,3	84,0
5	63,5	79,2	82,1	94,7	110,0
6	65,1	75,7	79,3	88,5	98,7
7	63,2	74,8	79,0	88,1	100,0
8	64,5	74,1	92,6	96,0	98,3

Fonte: GUIMARÃES et alii (1979)

Para obter a média aritmética aos 30 dias, basta calcular:

$$\bar{x} = \frac{76,2 + 81,5 + \ldots + 64,5}{8} = \frac{511,5}{8} = 63,94$$

Da mesma forma, para 34 dias obtém-se:

$$\bar{x} = \frac{95,5 + 90,0 + \ldots + 74,1}{8} = \frac{599,3}{8} = 74,91$$

As médias para as demais idades são obtidas de maneira idêntica. Essas médias, apresentadas na Tabela 4.6, mostram que o peso médio dos ratos aumenta com a idade.

Tabela 4.6

Peso médio de grupos de 8 ratos machos da raça Wistar segundo a idade, em dias

Idade				
30	34	38	42	46
63,94	74,91	81,65	95,05	105,89

4.5.2 - *Determine a mediana dos dados apresentados na Tabela 4.7.*

Tabela 4.7

Percentual de água em cérebro de cobaias machos com 90 dias de idade

80,06	68,86
68,97	79,90
79,85	79,91
79,87	79,55
79,86	79,25

Fonte: HOSSNE et alii (1990)

Para obter a mediana, os dados da Tabela 4.7 foram arranjados em ordem crescente na Tabela 4.8. Como o número de dados (10) é par, a mediana é a média aritmética dos dois valores que ocupam a posição central, ou seja, a mediana é:

$$\frac{79,85 + 79,86}{2} = 79,855.$$

Portanto, metade dos cérebros das cobaias da amostra tinha mais de 79,855% de água.

Tabela 4.8

Dados da Tabela 4.7 em ordem crescente

68,86	79,86
68,97	79,87
79,25	79,90
79,55	79,91
79,85	80,06

4.6 - Exercícios Propostos

4.6.1 - *Calcule a média dos dados apresentados na Tabela 4.9.*

Tabela 4.9

Taxa de glicose, em miligramas por 100 ml de sangue, em ratos machos da raça Wistar, com 20 dias de idade

100,0	97,5
100,0	85,0
97,5	85,0
80,0	80,0

Fonte: GUIMARÃES et alii (1979)

4.6.2 - *Determine a mediana dos dados apresentados na Tabela 4.10.*

Tabela 4.10

Peso corporal, em gramas, de ratos machos com 25 dias de idade

76	81
84	78
91	83
87	

Fonte: MUNHOZ et alii (1988)

4.6.3 - *Calcule o número médio de dentes cariados, para cada sexo, a partir dos dados apresentados na Tabela 4.11.*

Tabela 4.11

Escolares de 7 anos, segundo o número de dentes cariados e o sexo

Nº de dentes cariados	Sexo	
	Masculino	Feminino
0	16	13
1	2	5
2	3	3
3	2	2
4	2	2

Fonte: MOREIRA et alii (1985)

4.6.4 - *Determine a moda para os dados apresentados no Exercício 2.4.1 do Capítulo 2.*

CAPÍTULO
5

Medidas de Dispersão para uma Amostra

As medidas de tendência central, vistas no Capítulo 4, dão a abscissa do ponto em torno do qual os dados se distribuem. Estas medidas são tanto mais apropriadas para descrever a amostra quanto menor é a *dispersão* dos dados.

Para entender o que é dispersão, imagine que quatro alunos obtiveram, em cinco provas, as notas apresentadas na Tabela 5.1.

Tabela 5.1
Notas de quatro alunos em cinco provas

Aluno	Notas					Média
Antônio	5	5	5	5	5	5
João	6	4	5	4	6	5
José	10	5	5	5	0	5
Pedro	10	10	5	0	0	5

Todos os alunos obtiveram média igual a 5, mas a dispersão das notas em torno da média não é a mesma para todos os alunos. A Tabela 5.1 mostra claramente que:

a) As notas de Antônio não variaram (a dispersão é nula).
b) As notas de João variaram menos do que as notas de José (a dispersão das notas de João é menor do que a dispersão das notas de José).
c) As notas de Pedro variaram mais do que as notas de todos os outros (a dispersão das notas de Pedro é a maior).

Estas observações serão verificadas através das seguintes *medidas de dispersão*: amplitude, variância e desvio padrão.

5.1 - Amplitude

Por definição, *amplitude* é a diferença entre o maior e o menor dado observado. É fácil calcular a amplitude para os dados apresentados na Tabela 5.1. As notas de Antônio têm amplitude:

$$a = 5 - 5 = 0,$$

as notas de João têm amplitude:

$$a = 6 - 4 = 2,$$

as de José têm amplitude:

$$a = 10 - 0 = 10,$$

e as notas de Pedro têm amplitude:

$$a = 10 - 0 = 10.$$

A amplitude nem sempre capta certas diferenças. No caso das notas dos alunos, a amplitude mostra, acertadamente, que as notas de Antônio não variaram ($a = 0$) e que as notas de João variaram menos do que as notas de José ($a = 2$, no primeiro caso, e $a = 10$, no segundo caso). Entretanto a amplitude não mostra que as notas de Pedro variaram mais do que as notas de José ($a = 10$, nos dois casos).

A amplitude não mede bem a dispersão dos dados porque, em seu cálculo, usam-se apenas os valores extremos — e não todos os dados. De qualquer forma, a amplitude é muito usada, principalmente porque é fácil de calcular e fácil de interpretar.

5.2 - Variância

Os dados distribuem-se em torno da média. Então o grau de dispersão de um conjunto de dados pode ser medido pelos desvios em relação à média. *Desvio em relação à média* é a diferença entre cada dado e a média do conjunto. Por exemplo, se a média de idade numa família for 30 anos, a pessoa que tiver 54 anos terá um desvio em relação à média de:

$$54 - 30 = 24 \text{ anos.}$$

Como cada dado tem um desvio em relação à média, para julgar o grau de dispersão de uma amostra é preciso observar todos os desvios. Não se pode calcular a média dos desvios porque a soma é sempre igual a zero. Considere os seguintes dados:

$$0, 4, 6, 8 \text{ e } 7.$$

A média desses dados é:

$$\frac{0 + 4 + 6 + 8 + 7}{5} = \frac{25}{5} = 5$$

Os desvios em relação à média, representados por $x - \bar{x}$, são os seguintes:

$$\begin{aligned} 0 - 5 &= -5 \\ 4 - 5 &= -1 \\ 6 - 5 &= 1 \\ 8 - 5 &= 3 \\ 7 - 5 &= 2 \end{aligned}$$

Em conjunto, esses desvios mostram o grau de dispersão dos dados em torno da média. Mas a soma dos desvios é igual a zero, como é fácil verificar:

$$-5 - 1 + 1 + 3 + 2 = -6 + 6 = 0.$$

Qualquer que seja o conjunto de dados, a soma dos desvios é sempre igual a zero porque os valores positivos e negativos se anulam. Então, para medir a dispersão dos dados em torno da média, os estatísticos usam a *soma de quadrados dos desvios*. Como os quadrados de números negativos são positivos, toda soma de quadrados é positiva ou, no mínimo, nula (a soma dos quadrados dos desvios só é nula quando todos os desvios são iguais a zero).

É fácil calcular a soma de quadrados dos desvios. Veja o exemplo apresentado na Tabela 5.2 e verifique que a soma dos quadrados dos desvios é igual a 40.

Tabela 5.2

Cálculo da soma de quadrados dos desvios

Dados (x)	Desvios $(x-\bar{x})$	Quadrados dos desvios $(x-\bar{x})^2$
0	-5	25
4	-1	1
6	1	1
8	3	9
7	2	4
$\bar{x}=5$	$\Sigma(x-\bar{x})=0$	$\Sigma(x-\bar{x})^2=40$

A soma de quadrados dos desvios não é usada como medida de dispersão porque seu valor cresce com o número de dados. Para entender esta idéia, imagine dois grupos de pessoas. No primeiro grupo, as pessoas têm peso:

60, 70 e 80

e, no segundo grupo, as pessoas têm pesos:

60, 60, 70, 70, 80 e 80.

A dispersão dos dados em torno da média é a mesma, nos dois grupos. Mas compare as somas de quadrados dos desvios apresentados na última linha da Tabela 5.3. A soma é maior para o grupo II porque esse grupo tem mais dados. Então, mesmo que a dispersão se mantenha constante, a soma de quadrados dos desvios aumenta quando aumenta o número de dados.

Tabela 5.3

Cálculo da soma de quadrados dos desvios

Grupo I			Grupo II		
x	$(x-\bar{x})$	$(x-\bar{x})^2$	x	$(x-\bar{x})$	$(x-\bar{x})^2$
60	−10	100	60	−10	100
70	0	0	60	−10	100
80	10	100	70	0	0
			70	0	0
			80	10	100
			80	10	100
	0	200		0	400

Para medir a dispersão dos dados em torno da média usa-se, então, a *variância*, que leva em consideração o tamanho da amostra. A variância é

definida como a soma dos quadrados dos desvios dividida pelo tamanho da amostra, menos 1 ($n-1$). Os estatísticos chamam o valor $(n-1)$ de *graus de liberdade*. Portanto, a variância, que é indicada por s^2, é dada pela fórmula:

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

Desenvolvendo algebricamente a fórmula da variância, obtém-se:

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

Embora esta fórmula pareça, à primeira vista, difícil, ela na verdade facilita o trabalho de cálculo. Para conferir esta informação, calcule a variância dos dados 0, 4, 6, 8 e 7, usando esta fórmula. Os cálculos intermediários estão apresentados na Tabela 5.4.

Tabela 5.4

Cálculos intermediários para a obtenção da variância

x	x^2
0	0
4	16
6	36
8	64
7	49
$\sum x=25$	$\sum x^2=165$

Agora é fácil obter:

$$s^2 = \frac{165 - \frac{25^2}{5}}{4} = 10{,}0$$

Para entender que a variância mede a dispersão dos dados em torno da média, convém observar novamente as notas apresentadas na Tabela 5.1 e verificar que as variâncias são os valores dados na Tabela 5.5. Veja que a variância mede dispersão porque:

a) Para as notas de Antônio, que não variaram, $s^2 = 0$.
b) Para as notas de João, que variaram menos do que as notas de José, $s^2 = 1$, menor do que a variância das notas de José, que é $s^2 = 12{,}5$.
c) Para as notas de Pedro, que variaram mais do que todas as outras, a variância é $s^2 = 25$, maior do que todas as outras variâncias.

Tabela 5.5
Média e variância das notas de 4 alunos em 5 provas

Aluno	Notas					Média	Variância
Antônio	5	5	5	5	5	5	0
João	6	4	5	4	6	5	1
José	10	5	5	5	0	5	12,5
Pedro	10	10	5	0	0	5	25

5.3 - Desvio Padrão

Como medida de dispersão, a variância tem a desvantagem de apresentar unidade de medida igual ao quadrado da unidade de medida dos dados. Por exemplo, se ós dados estão em metros, a variância fica em metros ao quadrado.

Mas existe uma medida de dispersão que apresenta as propriedades da variância e tem a mesma unidade de medida dos dados. É o *desvio padrão*, definido como a raiz quadrada da variância, com sinal positivo. O desvio padrão é representado por s.

Para as notas do aluno José, cuja variância já foi calculada, tem-se o desvio padrão:

$$s = \sqrt{12,5} = 3,54.$$

5.4 - Coeficiente de Variação

O *coeficiente de variação* é a razão entre o desvio padrão e a média. O resultado é multiplicado por 100, para que o coeficiente de variação seja dado em porcentagem. Então:

$$CV = \frac{s}{\bar{x}} \cdot 100.$$

Para entender como se interpreta o coeficiente de variação, imagine dois grupos de pessoas. No primeiro grupo, as pessoas têm idades

3, 1 e 5

e no segundo grupo as pessoas têm idades

55, 57 e 53.

No primeiro grupo, a média de idade é 3 anos e, no segundo grupo, a média de idade é 55 anos. Nos dois grupos a dispersão dos dados é a mesma. Ambos têm variância $s^2 = 4$. Mas as diferenças de dois anos são muito mais importantes no primeiro grupo, que tem média 3, do que no segundo grupo, que tem média 55. Agora, veja os coeficientes de variação. No primeiro grupo, o coeficiente de variação é:

$$CV = \frac{2}{3} \cdot 100 = 66{,}67\%$$

e, no segundo grupo, o coeficiente de variação é:

$$CV = \frac{2}{55} \cdot 100 = 3{,}64\% .$$

Um coeficiente de variação igual a 66,67% indica que a dispersão dos dados em relação à média é muito grande, ou seja, a *dispersão relativa* é alta. Já um coeficiente de variação de 3,64% indica que a dispersão dos dados em relação à média é pequena. Em outras palavras, diferenças de 2 anos são relativamente mais importantes no primeiro grupo, que tem média 3 (o coeficiente de variação é 66,67%) do que no segundo grupo, que tem média 55 (o coeficiente de variação é 3,64%). Então o coeficiente de variação mede *dispersão em relação à média*.

5.5 - Exercícios Resolvidos

5.5.1 - *Calcule a variância e o desvio padrão dos dados apresentados na Tabela 4.5 do Capítulo 4, em cada idade. Comente o resultado.*

A variância é dada pela fórmula:

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

Usando uma máquina de calcular, é fácil obter:

a) Para 30 dias de idade:

$$\sum x^2 = 33\,629{,}69;\ \sum x = 511{,}5;\ (\sum x)^2 = 261\,632{,}25$$

$$s^2 = \frac{925{,}66}{7} = 132{,}24$$

b) Para 34 dias de idade:

$$\sum x^2 = 46\,409{,}23;\ \sum x = 599{,}3;\ (\sum x)^2 = 359\,160{,}49$$

$$s^2 = \frac{1514,17}{7} = 216,31$$

c) Para 38 dias de idade:

$$\Sigma x^2 = 55\,114,28;\ \Sigma x = 653,2;\ (\Sigma x)^2 = 426\,670,24$$

$$s^2 = \frac{1780,50}{7} = 254,36$$

d) Para 42 dias de idade:

$$\Sigma x^2 = 75\,124,18;\ \Sigma x = 760,4;\ (\Sigma x)^2 = 578\,208,16$$

$$s^2 = \frac{2848,16}{7} = 406,88$$

e) Para 46 dias de idade:

$$\Sigma x^2 = 92\,504,27;\ \Sigma x = 847,1;\ (\Sigma x)^2 = 717\,578,41$$

$$s^2 = \frac{2806,97}{7} = 401,00$$

Para calcular o desvio padrão basta extrair a raiz quadrada da variância. Os valores dos desvios padrões estão apresentados na Tabela 5.6. É fácil ver que os desvios padrões aumentam com a idade. Portanto, a dispersão dos dados em torno da média aumenta com a idade.

Tabela 5.6

Desvio padrão do peso de grupos de 8 ratos machos da raça Wistar, segundo a idade, em dias

Idade				
30	34	38	42	46
11,50	14,71	15,95	20,17	20,02

5.5.2 - *Calcule a média, o desvio padrão e o coeficiente de variação dos dados apresentados na Tabela 5.7. Comente os resultados.*

Usando uma máquina de calcular, é fácil obter:

a) Para peso:

$$\bar{x} = \frac{205,8}{10} = 20,58$$

$$s^2 = \frac{4364,04 - 4235,36}{9} = 14,30$$

$$s = 3{,}78$$

$$CV = \frac{3{,}78}{20{,}58} \cdot 100 = 18{,}37\%$$

b) Para comprimento:

$$\bar{x} = \frac{1023}{10} = 102{,}3$$

$$s^2 = \frac{104865{,}0 - 104652{,}9}{9} = 23{,}57$$

$$s = 4{,}85$$

$$CV = \frac{4{,}85}{102{,}3} \cdot 100 = 4{,}74\%$$

Tabela 5.7

Peso, em quilogramas, e comprimento, em centímetros, de 10 cães

Peso	Comprimento
23,0	104
22,7	107
21,2	103
21,5	105
17,0	100
28,4	104
19,0	108
14,5	91
19,0	102
19,5	99

Fonte: ARAÚJO e HOSSNE (1977)

Não se podem comparar desvios padrões de peso e comprimento porque as unidades de medida são diferentes. Mas os coeficientes de variação podem ser comparados porque são adimensionais. Então é fácil ver que a dispersão relativa dos dados de peso (CV = 18,37%) é maior do que a dispersão relativa dos dados de comprimento (CV = 4,74%). Isto significa que os dados de comprimento variam menos em relação à média do que os dados de peso.

5.6 - Exercícios Propostos

5.6.1 - *Calcule a variância, o desvio padrão e o coeficiente de variação dos dados apresentados no Exercício 4.6.1 do Capítulo 4.*

5.6.2 - *Calcule a amplitude dos dados apresentados no Exercício 4.6.2 do Capítulo 4.*

5.6.3 - *Calcule a média e o desvio padrão dos dados apresentados na Tabela 5.8, para cada sexo.*

Tabela 5.8

Comprimento, em centímetros, de cobaias de 90 dias, segundo o sexo

Sexo	
Masculino	Feminino
25,5	27,0
26,0	27,0
26,5	27,0
25,0	27,0
26,0	26,0
25,0	27,0
24,0	27,5
25,0	27,0
25,5	28,0
26,0	26,0

Fonte: HOSSNE et alii (1990)

5.6.4 - *Calcule a média e o desvio padrão dos dados apresentados em cada coluna da Tabela 5.9.*

Tabela 5.9

Peso fresco, em gramas, de pulmões de cobaias machos de 90 dias de idade

Pulmão	
Direito	Esquerdo
1,66	1,48
2,15	1,58
2,03	1,59
2,35	1,92
1,90	1,55

Fonte: HOSSNE et alii (1990)

CAPÍTULO
6

Noções sobre Correlação

Existem situações nas quais interessa estudar o comportamento conjunto de duas variáveis. Por exemplo, dados peso e estatura de pessoas, pode haver interesse em estabelecer em que medida aumenta o peso, quando a estatura aumenta.

O comportamento conjunto de duas variáveis quantitativas pode ser observado através de um gráfico, denominado diagrama de dispersão, e medido através do coeficiente de correlação. Estes dois procedimentos serão vistos neste Capítulo.

6.1 - Diagrama de Dispersão

Para desenhar um *diagrama de dispersão*, primeiro se traça o sistema de eixos cartesianos. Depois se representa uma das variáveis no eixo das abscissas e a outra variável no eixo das ordenadas. Colocam-se, então, os valores das variáveis sobre os respectivos eixos e marca-se um ponto para cada par de valores.

Observe os dados de comprimento e peso de cães, apresentados na Tabela 6.1 e em diagrama de dispersão na Figura 6.1. A variável comprimento foi representada no eixo das abscissas e a variável peso foi representada no eixo das ordenadas. O diagrama mostra que comprimento e peso de cães crescem no mesmo sentido.

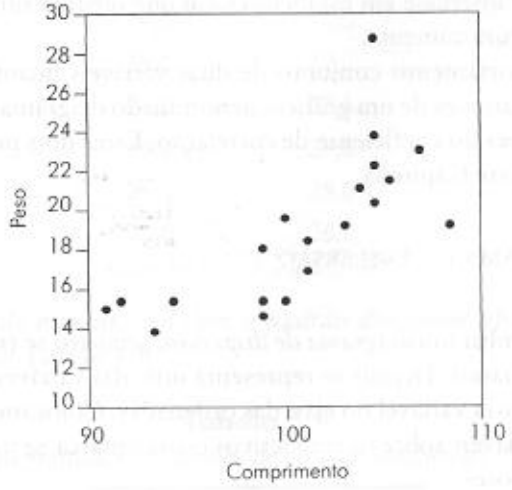
Tabela 6.1

Comprimento, em centímetros, e peso, em quilogramas, de cães

Comprimento	Peso	Comprimento	Peso
104	23,5	98	15,0
107	22,7	95	14,9
103	21,1	92	15,1
105	21,5	104	22,2
100	17,0	94	13,6
104	28,5	99	16,1
108	19,0	98	18,0
91	14,5	98	16,0
102	19,0	104	20,0
99	19,5	100	18,3

Fonte: ARAÚJO e HOSSNE (1977)

Figura 6.1 *Comprimento, em centímetros, e peso, em quilogramas, de cães*



6.2 - Correlação Positiva e Correlação Negativa

Se as variáveis X e Y crescem no mesmo sentido, isto é, se quando X cresce Y em média também cresce, diz-se que as duas variáveis têm *correlação positiva*. Então, peso e comprimento de cães têm correlação positiva porque, quando uma das variáveis cresce, a outra, em média, também cresce.

Se as variáveis X e Y variam em sentidos contrários, isto é, se quando X cresce, Y em média decresce, diz-se que as duas variáveis têm *correlação negativa*. Observe os dados apresentados na Tabela 6.2 e mostrados em diagrama de dispersão na Figura 6.2. É fácil ver que consumo individual diário de proteínas de origem animal e coeficiente de natalidade variam em sentidos contrários. Então essas variáveis têm correlação negativa.

Tabela 6.2

Consumo individual diário de proteínas de origem animal, em gramas, e coeficiente de natalidade, em 14 países

País	Consumo individual diário de proteínas	Coeficiente de natalidade
Formosa	4,7	45,6
Malásia	7,5	39,7
Índia	8,7	33,0
Japão	9,7	27,0
Iugoslávia	11,2	25,9
Grécia	15,2	23,5
Itália	15,2	23,4
Bulgária	16,8	22,2
Alemanha	37,3	20,0
Irlanda	46,7	19,1
Dinamarca	56,1	18,3
Austrália	59,9	18,0
Estados Unidos	61,4	17,9
Suécia	62,6	15,0

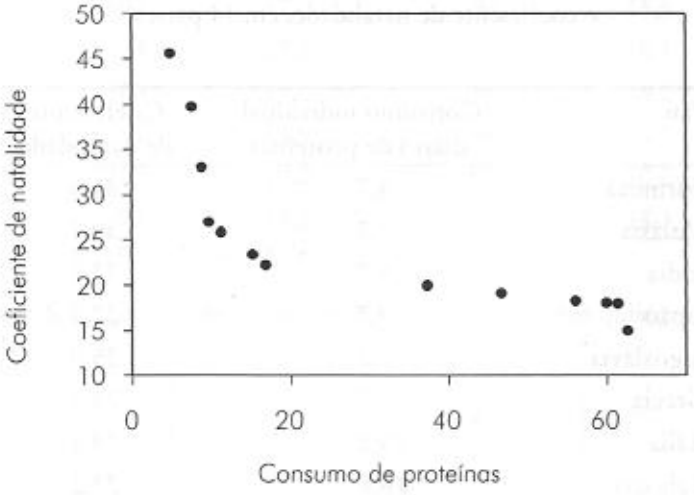
Fonte: CASTRO (1961)

É importante deixar claro, aqui, que uma correlação positiva entre duas variáveis mostra apenas que essas variáveis crescem no mesmo sentido. A correlação positiva não indica que aumentos sucessivos em uma das variáveis *causam* aumentos sucessivos na outra variável. Da mesma forma, uma correlação negativa entre duas variáveis mostra apenas que elas variam em sentidos contrários. A correlação negativa não indica que acréscimos em uma das variáveis *causam* decréscimos na outra variável.

Observe os dados apresentados na Tabela 6.2. Existe correlação negativa entre consumo individual diário de proteínas de origem animal

e coeficiente de natalidade, mas isso não significa que um aumento no consumo de proteínas de origem animal causa redução de fertilidade. A correlação negativa entre as duas variáveis talvez seja explicada pela qualidade de vida. É razoável admitir que a melhoria na qualidade de vida de um país determina tanto um aumento no consumo médio de proteínas como uma diminuição no coeficiente de natalidade.

Figura 6.2 *Consumo individual diário de proteínas de origem animal e coeficiente de natalidade, em 14 países*



6.3 - Coeficiente de Correlação

Existe uma medida para o grau de correlação entre duas variáveis. Essa medida é o *coeficiente de correlação de Pearson*, que se representa por r e é definido pela fórmula:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

Para entender como se aplica esta fórmula, primeiro observe os exemplos — um de correlação positiva e outro de correlação negativa — apresentados na Tabela 6.3 e na Figura 6.3.

Para se obter o coeficiente de correlação — no caso do exemplo de correlação positiva da Tabela 6.3 — foram feitos os cálculos intermediários apresentados na Tabela 6.4.

Tabela 6.3
Um exemplo de correlação positiva e um exemplo de correlação negativa

Correlação			
Positiva		Negativa	
x	y	x	y
1	1	1	7
2	2	2	4
3	4	3	4
4	5	4	3
5	8	5	1

Figura 6.3 *Exemplos de correlação positiva e negativa*

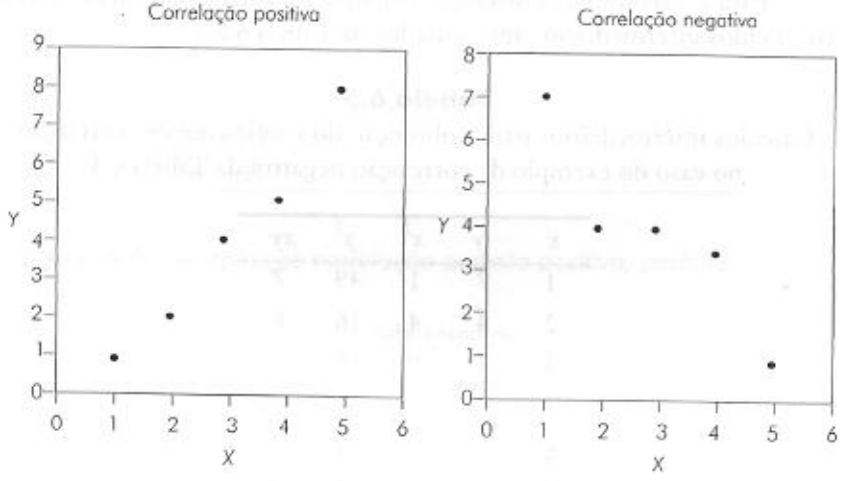


Tabela 6.4
Cálculos intermediários para a obtenção do coeficiente de correlação no caso do exemplo de correlação positiva da Tabela 6.3

x	y	x^2	y^2	xy
1	1	1	1	1
2	2	4	4	4
3	4	9	16	12
4	5	16	25	20
5	8	25	64	40
15	20	55	110	77

Com os valores apresentados na Tabela 6.4, obtém-se:

$$r = \frac{77 - \frac{15 \cdot 20}{5}}{\sqrt{\left(55 - \frac{15^2}{5}\right)\left(110 - \frac{20^2}{5}\right)}}$$

$$= \frac{77 - 60}{\sqrt{(55 - 45)\,(110 - 80)}}$$

$$= \frac{17}{\sqrt{300}}$$

$$= 0{,}98$$

Para o exemplo de correlação negativa da Tabela 6.3, foram feitos os cálculos intermediários apresentados na Tabela 6.5.

Tabela 6.5

Cálculos intermediários para a obtenção do coeficiente de correlação no caso do exemplo de correlação negativa da Tabela 6.3

x	y	x^2	y^2	xy
1	7	1	49	7
2	4	4	16	8
3	4	9	16	12
4	3	16	9	12
5	1	25	1	5
15	19	55	91	44

$$r = \frac{44 - \frac{15 \cdot 19}{5}}{\sqrt{\left(55 - \frac{15^2}{5}\right)\left(91 - \frac{19^2}{5}\right)}}$$

$$= \frac{44 - 57}{\sqrt{(55 - 45)\,(91 - 72{,}2)}}$$

$$= -\frac{13}{\sqrt{188}}$$

$$= -0{,}95$$

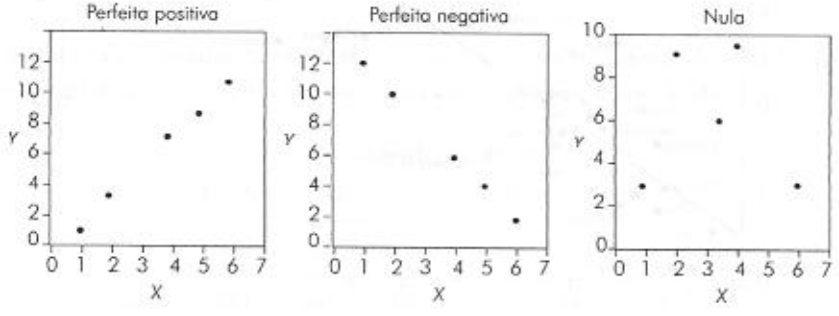
O coeficiente de correlação varia entre –1 e +1, inclusive, isto é, $-1 \leq r \leq +1$. Se r assume o valor 1, diz-se que as duas variáveis têm *correlação perfeita positiva* e se r assume o valor –1, diz-se que as duas variáveis têm *correlação perfeita negativa*. Se r assume o valor zero, não existe correlação entre as duas variáveis (a correlação é nula). Observe os exemplos apresentados na Tabela 6.6 e, em diagramas de dispersão, na Figura 6.4.

Tabela 6.6

Um exemplo de correlação perfeita positiva, um exemplo de correlação perfeita negativa e um exemplo de correlação nula

Correlação					
Perfeita positiva		Perfeita negativa		Nula	
x	y	x	y	x	y
1	1	1	12	1	3
2	3	2	10	2	9
4	7	4	6	4	6
5	9	5	4	5	9
6	11	6	2	6	3

Figura 6.4 *Exemplos de correlação perfeita positiva, perfeita negativa e nula*



6.4 - Exercícios Resolvidos

6.4.1 - *Calcule o coeficiente de correlação para os dados apresentados na Tabela 6.2.*

Primeiro é preciso obter:

$\sum xy = 8\,217{,}85$ $\quad \sum x \sum y = 143\,971{,}8$

$\sum x = 413{,}0$ $\quad \sum x^2 = 19\,114{,}0$

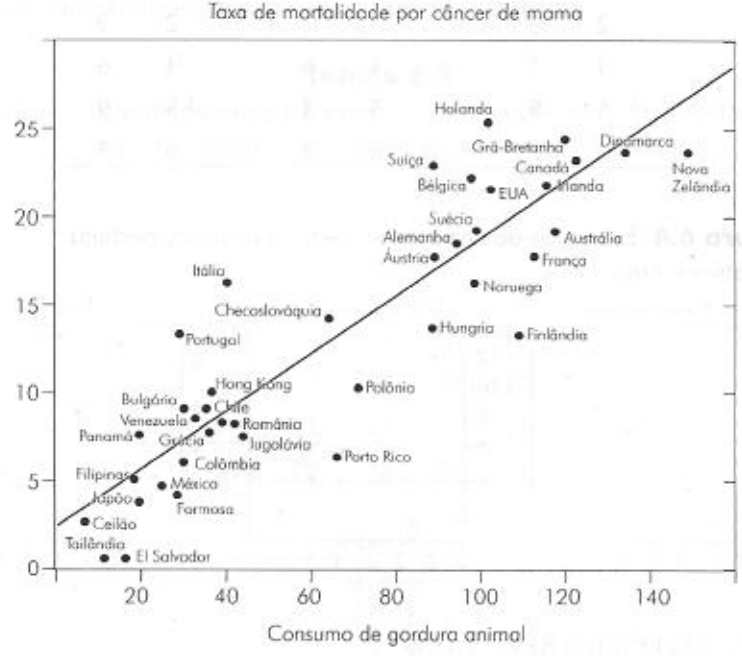
$\sum y = 348{,}6$ $\quad \sum y^2 = 9\,706{,}02$

Aplicando a fórmula, vem:
$r = -0{,}775$

Portanto, os dados de consumo individual diário de proteína de origem animal e os coeficientes de natalidade apresentados na Tabela 6.2 têm correlação negativa igual a – 0,775.

6.4.2 - *Os diagramas de dispersão apresentados nas Figuras 6.5 e 6.6 mostram a relação entre o consumo per capita de gordura animal e de gordura vegetal, em gramas/dia, e a taxa de mortalidade por câncer de mama, em 39 países. Parece existir forte relação entre consumo de gordura animal e câncer de mama. Isso significa que a gordura animal causa câncer de mama?*

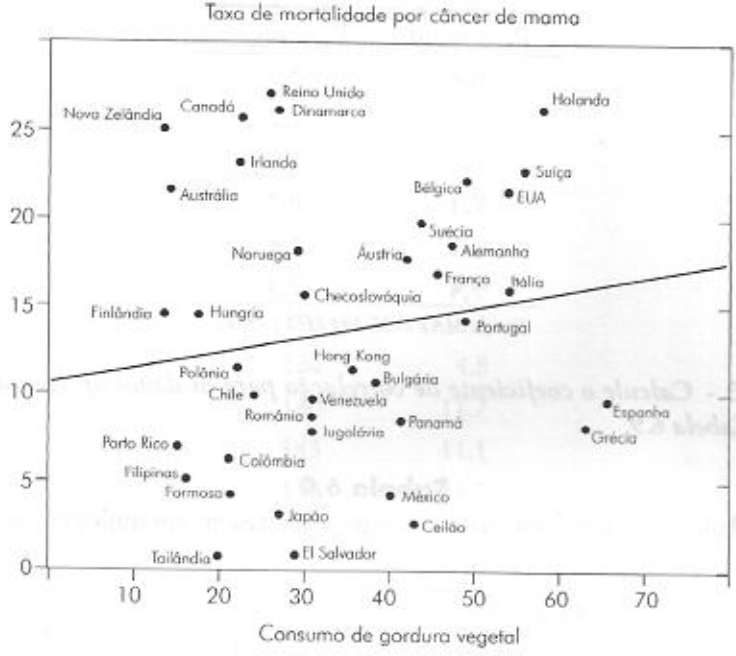
Figura 6.5 *Consumo diário per capita de gordura animal e taxa de mortalidade por câncer de mama, em 39 países*



Fonte: CAROLL (1975)

Pode não existir relação de causa e efeito entre consumo de gordura animal e câncer de mama. Sabe-se, por exemplo, que o risco de câncer de mama está relacionado com outras variáveis como renda, uso de automóveis e número de aparelhos de televisão na residência. Então o risco de câncer de mama talvez esteja relacionado com vários indicadores de vida rica e sedentária, e não apenas com o consumo de gordura animal.

Figura 6.6 *Consumo diário per capita de gordura vegetal e taxa de mortalidade por câncer de mama, em 39 países*



Fonte: CAROLL (1975)

6.5 - Exercícios Propostos

6.5.1 - *Faça um diagrama de dispersão e calcule o coeficiente de correlação para os dados apresentados na Tabela 6.7. Discuta o resultado.*

Tabela 6.7

Dados relativos a duas variáveis X e Y

X	Y
3	2
5	2
4	7
2	7
1	2

6.5.2 - *Calcule o coeficiente de correlação para os dados apresentados na Tabela 6.8.*

Tabela 6.8

Peso úmido e peso seco, em gramas, de lóbulos hepáticos de ratos

Peso úmido	Peso seco
6,7	2,0
7,7	2,2
6,5	2,0
7,4	2,2
6,1	1,9
7,4	2,3

Fonte: MATTOS FILHO (1976)

6.5.3 - *Calcule o coeficiente de correlação para os dados apresentados na Tabela 6.9.*

Tabela 6.9

Idade gestacional, em semanas, e peso ao nascer, em quilogramas, de recém-nascidos

Idade gestacional	Peso ao nascer
28	1,25
32	1,25
35	1,75
38	2,25
39	3,25
41	3,25
42	4,25

6.5.4 - *Em um trabalho sobre acumulação de placa dental em pacientes jovens, foi obtido tanto um índice clínico para medir a quantidade de placa como o peso seco das placas, em miligramas. Os dados estão na Tabela 6.10. Construa um diagrama de dispersão. Você acha que existe correlação entre as medidas?*

Tabela 6.10

Índice clínico e peso seco, em miligramas, das placas dentais em 10 pacientes

Índice clínico	Peso seco
25	2,7
45	2,7
60	3,5
68	3,7
80	5,8
100	5,1
120	4,8
140	11,7
143	11,1
148	14,2

Fonte: ASHLEY et alii (1984)

CAPÍTULO
7

Noções sobre Regressão

Muitas vezes interessa estudar o comportamento conjunto de duas variáveis, como mostra o Capítulo 6. Outras vezes interessa estudar como uma variável varia em função de outra. Por exemplo, considere a questão de idade e peso das crianças. Sempre existe interesse em estudar como o peso varia em função da idade.

Quando se estuda a variação de uma variável Y em função de uma variável X, diz-se que Y é a *variável dependente* e que X é a *variável explanatória*. No caso do exemplo, sabe-se que o peso das crianças varia em função da idade. Então peso é a variável dependente e idade é a variável explanatória.

7.1 - Gráfico de Linhas

É possível observar a variação de uma variável em função de outra, através do *gráfico de linhas*. Para fazer o gráfico de linhas, primeiro se traça o sistema de eixos cartesianos. Depois se representa a variável explanatória no eixo das abscissas e a variável dependente no eixo das ordenadas. Colocam-se então os valores das variáveis sobre os respectivos eixos e marca-se um ponto para cada par de valores. Finalmente, considerando a seqüência de valores crescentes de X, unem-se os pontos por segmentos de reta. Os dados da Tabela 7.1 estão apresentados em gráfico de linhas na Figura 7.1.

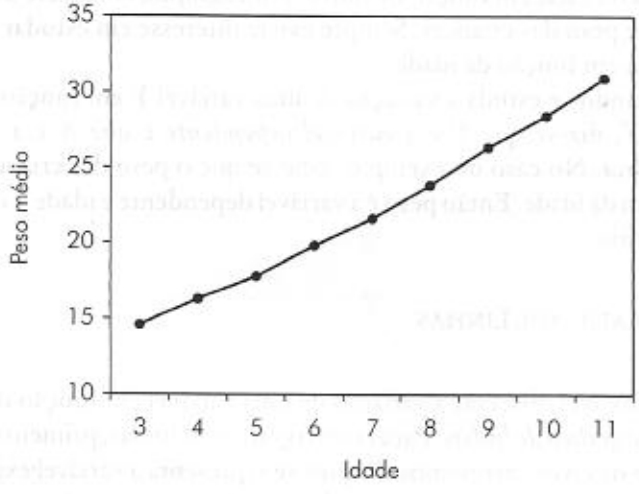
Tabela 7.1

Peso médio, em quilogramas, de indivíduos do sexo masculino, segundo a idade, no Distrito Federal

Idade	Peso médio	Idade	Peso médio
3	14,6	12	34,2
4	16,3	13	38,7
5	17,8	14	43,4
6	19,8	15	49,7
7	21,6	16	52,7
8	23,8	17	57,3
9	26,3	18	58,1
10	28,4	19	59,4
11	30,9		

Fonte: IBGE (1978)

Figura 7.1 *Peso médio, em quilogramas, de indivíduos do sexo masculino, segundo a idade, no Distrito Federal. IBGE, 1978*



7.2 - Reta de Regressão

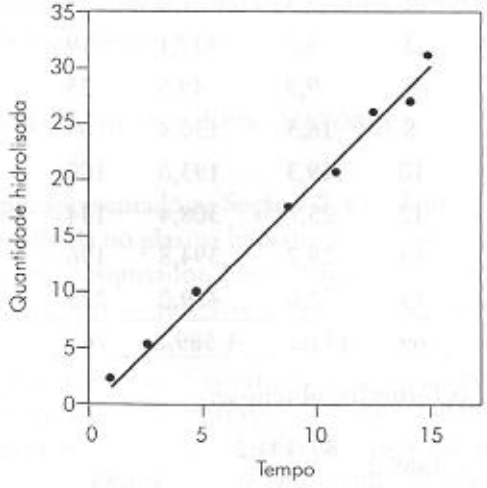
A idéia de *regressão* fica bem entendida através de um exemplo. Observe os dados apresentados na Tabela 7.2. É fácil ver que a quantidade de procaína (anestésico local) hidrolisada no plasma humano varia em função do tempo decorrido após sua administração.

Tabela 7.2
Quantidade de procaína hidrolisada, em 10 moles/litro, no plasma humano, em função do tempo decorrido após sua administração

Tempo (minutos)	Quantidade hidrolisada
2	3,5
3	5,7
5	9,9
8	16,3
10	19,3
12	25,7
14	28,2
15	32,6

Fonte: AVEN e FOLDES (1951)

Figura 7.2 *Quantidade de procaína hidrolisada, em 10 moles/litro, no plasma humano, em função do tempo decorrido após sua administração*



Os dados da Tabela 7.2 estão apresentados em diagrama de dispersão na Figura 7.2. Note que os pontos estão praticamente sobre uma reta. Logo, a variação da quantidade de procaína hidrolisada no plasma humano em função do tempo decorrido após sua administração pode ser descrita através de uma reta que, em estatística, recebe o nome de *reta de regressão*.

Para *ajustar uma regressão linear simples* (isto é, a equação de uma reta) aos dados apresentados na Tabela 7.2, é preciso obter os coeficientes linear e angular da reta.

O coeficiente angular — que dá a inclinação da reta — é representado por b e é obtido através da fórmula:

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

O coeficiente linear — que é a ordenada do ponto em que a reta corta o eixo das ordenadas — é representado por a e é obtido através da fórmula:

$$a = \bar{y} - b\bar{x},$$

onde $\bar{y}$ e $\bar{x}$ são as médias de Y e X, respectivamente.

Reveja, agora, os dados apresentados na Tabela 7.2 e faça os cálculos intermediários apresentados na Tabela 7.3.

Tabela 7.3

Cálculos intermediários para a obtenção de a e de b

x	y	xy	x^2
2	3,5	7,0	4
3	5,7	17,1	9
5	9,9	49,5	25
8	16,3	130,4	64
10	19,3	193,0	100
12	25,7	308,4	144
14	28,2	394,8	196
15	32,6	489,0	225
69	141,2	1 589,2	767

Aplicando as fórmulas, obtém-se:

$$b = \frac{1589{,}2 - \frac{69 \cdot 141{,}2}{8}}{767 - \frac{69^2}{8}} = \frac{371{,}35}{171{,}875} = 2{,}16$$

$$a = \frac{141{,}2}{8} - 2{,}16\,\frac{69}{8} = -0{,}98$$

Para traçar a *reta de regressão* é preciso dar valores arbitrários para X e depois calcular os valores de Y. Indicam-se os valores calculados de Y por $\hat{Y}$. Fazendo $X = 5$, tem-se que:

$$\hat{Y} = -0{,}98 + 2{,}16 \cdot 5 = 9{,}82$$

e fazendo $X = 15$, tem-se que:

$$\hat{Y} = -0{,}98 + 2{,}16 \cdot 15 = 31{,}42.$$

Os dois pares de valores ($X = 5$ e $\hat{Y} = 9{,}82$) e ($X = 15$ e $\hat{Y} = 31{,}42$), colocados no gráfico da Figura 7.2, permitem traçar a reta de regressão ali apresentada, cuja *equação* é:

$$\hat{Y} = -0{,}98 + 2{,}16\,X.$$

A equação da reta de regressão permite calcular os valores de $\hat{Y}$ para quaisquer valores de X dentro do intervalo estudado, mesmo que esses valores não existam na amostra. Observe os dados apresentados na Tabela 7.2. Não existe o valor $X = 13$, mas, para calcular $\hat{Y}$, basta fazer:

$$\hat{Y} = -0{,}98 + 2{,}16 \cdot 13 = 27{,}10$$

O valor $\hat{Y} = 27{,}10$ é uma *previsão*, feita com base na equação da reta de regressão, para a quantidade de procaína que estaria hidrolisada 13 minutos após sua administração.

7.3 - Escolha da Variável Explanatória

No exemplo apresentado na Seção 7.2, é evidente que a quantidade de procaína hidrolisada no plasma humano foi medida em tempos previamente fixados pelo pesquisador. Nessas situações — em que os valores de X são fixados *a priori* — ajusta-se a regressão de Y contra X.

Mas nem sempre os valores de X são fixados *a priori*. Então tanto se pode ajustar a regressão de Y contra X, como a regressão de X contra Y. Para escolher entre as duas regressões, é razoável identificar a variável que deve ser prevista, conhecido o valor da outra variável. Ajusta-se a regressão de Y contra X toda vez que se pretende estudar a variação de Y (prever Y), em função da variação de X.

Observe os dados apresentados na Tabela 7.4. É razoável estudar a variação da pressão arterial (Y) em função do peso de cães (X). Então se deve ajustar uma regressão de Y contra X.

Tabela 7.4

Pressão arterial (P.A.), em milímetros de mercúrio, e peso de cães, em quilogramas

P.A.	Peso	P.A.	Peso	P.A.	Peso
130,0	23,0	135,0	23,8	90,5	16,0
107,5	22,7	125,0	22,0	115,5	20,0
135,0	21,2	110,0	18,7	113,0	18,3
100,0	21,5	102,0	19,5	116,0	22,3
134,5	17,0	121,5	28,0	143,0	24,0
121,5	28,4	111,5	15,0	104,5	15,8
107,5	19,0	107,5	18,8	102,5	16,0
105,0	14,5	127,5	20,5	107,5	15,0
125,0	19,0	104,5	15,0	125,5	16,0
130,0	19,5	102,5	14,9	93,0	22,5

Fonte: ARAÚJO e HOSSNE (1977)

Foram calculados:

$$b = \frac{68316{,}15 - \dfrac{587{,}9 \cdot 3454{,}0}{30}}{11931{,}19 - \dfrac{(587{,}9)^2}{30}} = 1{,}534$$

e

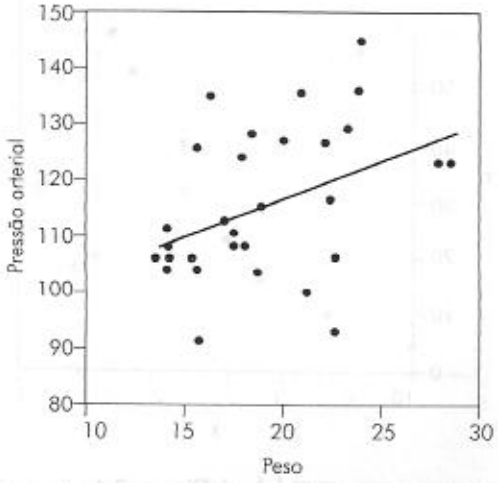
$$a = 115{,}13 - 1{,}534 \cdot 19{,}597 = 85{,}07.$$

A reta de regressão

$$\hat{Y} = 85{,}07 + 1{,}534X,$$

apresentada na Figura 7.3, permite prever a pressão arterial de cães em função do peso. Mas convém observar a Figura 7.3 com atenção. Note que os pontos estão muito dispersos em torno da reta. Isso significa que a previsão da pressão arterial de um cão, com base em seu peso, tem grande margem de erro. Mas é fácil ver a tendência de ocorrer aumento de pressão arterial quando aumenta o peso.

Figura 7.3 *Reta de regressão para pressão arterial em função do peso, em cães*



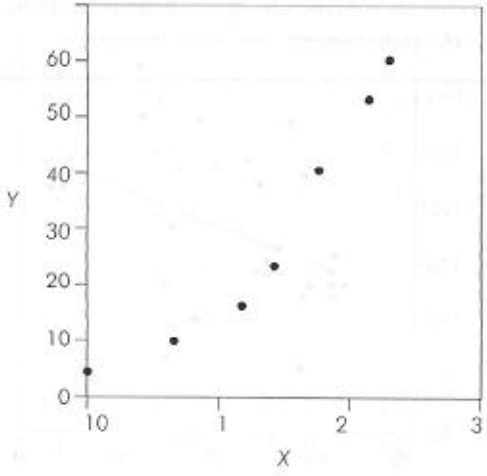
7.4 - TRANSFORMAÇÃO DE VARIÁVEIS

Existem situações em que os pares de valores das variáveis X e Y, apresentados em diagrama de dispersão, não se distribuem em torno de uma reta. Observe os dados da Tabela 7.5, apresentados em diagrama de dispersão na Figura 7.4. Note que os pontos estão dispersos em torno de uma curva.

Tabela 7.5
Valores de duas variáveis quaisquer X e Y

X	Y
0	4,0
0,6	8,0
1,2	15,0
1,5	22,6
1,8	36,4
2,1	45,3
2,4	60,0

Figura 7.4 *Diagrama de dispersão*



Como os pontos apresentados na Figura 7.4 não estão em torno de uma reta, pode-se experimentar *transformar* a variável Y. Em termos práticos, pode-se fazer um diagrama de dispersão com o logaritmo decimal de Y, em lugar de Y. Os valores de X e dos logaritmos decimais de Y estão apresentados na Tabela 7.6 e na Figura 7.5. A definição de logaritmo é dada na Seção 15.4 do Capítulo 15.

Tabela 7.6
Valores de X e valores dos logaritmos de Y

X	log$Y$
0	0,602
0,6	0,903
1,2	1,176
1,5	1,354
1,8	1,561
2,1	1,656
2,4	1,778

Os pontos relativos às variáveis X e log$Y$ estão praticamente sobre uma reta. É possível, então, ajustar uma regressão linear simples de log$Y$ contra X. Para calcular a e b, são necessários os cálculos intermediários apresentados na Tabela 7.7.

Figura 7.5 *Diagrama de dispersão*

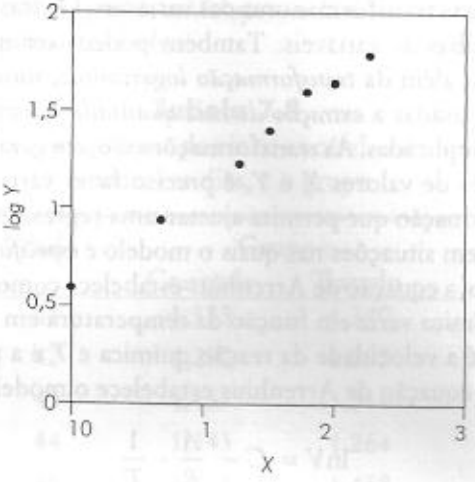


Tabela 7.7
Cálculos intermediários para a obtenção de a e b

X	log$Y$	Xlog$Y$	X^2
0	0,602	0	0
0,6	0,903	0,5418	0,36
1,2	1,176	1,4112	1,44
1,5	1,354	2,0310	2,25
1,8	1,561	2,8098	3,24
2,1	1,656	3,4776	4,41
2,4	1,778	4,2672	5,76
9,6	9,030	14,5386	17,46

Com base nos cálculos apresentados na Tabela 7.7, é fácil obter:

$$b = \frac{14{,}5386 - \dfrac{9{,}6 \cdot 9{,}030}{7}}{17{,}46 - \dfrac{(9{,}6)^2}{7}} = 0{,}502$$

$$a = \frac{9{,}030}{7} - 0{,}502 \cdot \frac{9{,}6}{7} = 0{,}602$$

Portanto, a equação de reta de regressão de log$Y$ contra X é:

$$\log\hat{Y} = 0{,}602 + 0{,}502X$$

Para que uma regressão linear simples possa ser ajustada aos dados, muitas vezes basta transformar uma das variáveis. Outras vezes, é preciso transformar ambas as variáveis. Também podem ser utilizadas outras transformações, além da *transformação logarítmica*, mostrada aqui. Assim, são muito usadas a *extração de raiz quadrada* e a *inversão*, além de outras, mais complicadas. As transformações são, em geral, *empíricas*, isto é, dados n pares de valores X e Y, é preciso fazer várias tentativas, até achar a transformação que permita ajustar uma regressão linear simples.

Mas existem situações nas quais o modelo é *especificado* teoricamente. Por exemplo, a equação de Arrenhius estabelece como a velocidade de uma reação química varia em função da temperatura em que se processa a reação. Se V é a velocidade da reação química e T é a temperatura em graus Kelvin, a equação de Arrenhius estabelece o modelo:

$$\ln V = C - \frac{A}{R} \cdot \frac{1}{T}$$

onde $\ln V$ é o logaritmo neperiano da velocidade da reação química à temperatura T, e R é uma constante (1,987 cal/grau/mol).

Para ajustar a equação de Arrenhius a pares de dados relativos à velocidade da reação e à temperatura, é preciso primeiro calcular os valores das variáveis transformadas, isto é, o *logaritmo neperiano da velocidade* e o *inverso da temperatura*. Depois se ajusta uma regressão linear simples do logaritmo neperiano de V contra o inverso de T, isto é:

$$\ln V = a + b \frac{1}{T}$$

Então, $C = a$ e $A = -Rb$.

Cabe esclarecer, aqui, que nem sempre é possível ajustar uma regressão linear simples. Existem situações que exigem o uso de modelos matemáticos mais complexos. É o caso, por exemplo, dos dados de crescimento, que, apresentados em gráfico, dão origem às *curvas de crescimento*. Mas este assunto não será tratado aqui.

7.5 - Exercícios Resolvidos

7.5.1 - PERALTA et alii (1976) separaram 100 ratos de 40 dias de idade em dois grupos de 50: um grupo (controle) recebeu água à vontade, enquanto o outro grupo (tratado) foi privado de água. Depois, em intervalos de 24 horas, sacrificaram cinco ratos de cada grupo. Os encéfalos desses ratos foram removidos e pesados. Como no início do experimento todos os ratos tinham 40 dias, os primeiros dez ratos sacrificados tinham 41 dias no dia do sacrifício, os dez seguintes tinham 42, e assim por diante, até os dez últimos,

que tinham 50. As médias dos pesos dos cinco encéfalos, segundo a idade no dia do sacrifício e o grupo, estão apresentados na Tabela 7.8. Mostre esses dados em gráficos de linhas.

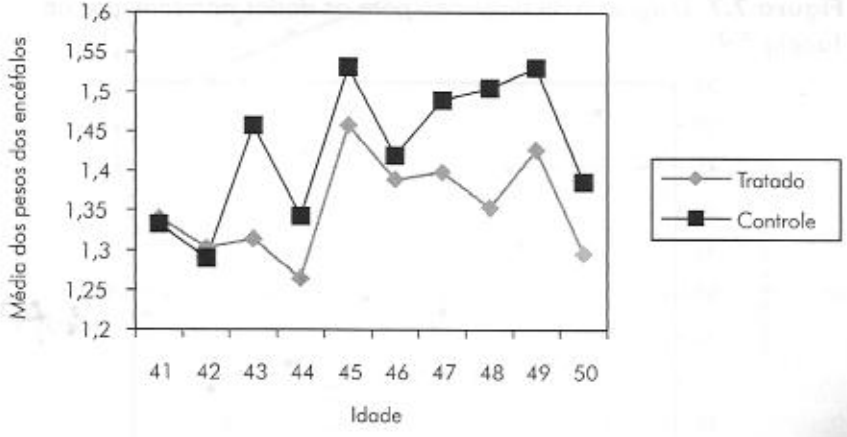
Tabela 7.8

Média dos pesos, em gramas, de cinco encéfalos de ratos segundo a idade, em dias, e o grupo

Idade	Grupo	
	Controle	Tratado
41	1,333	1,340
42	1,290	1,303
43	1,457	1,314
44	1,343	1,264
45	1,531	1,458
46	1,420	1,389
47	1,490	1,399
48	1,505	1,354
49	1,530	1,427
50	1,386	1,296

Fonte: PERALTA et alii (1976)

Figura 7.6 *Média dos pesos, em gramas, de cinco encéfalos de ratos em função da idade, em dias, para dois grupos*



A Figura 7.6 mostra que a média dos pesos de cinco encéfalos de ratos cresceu em função da idade e que, nos ratos privados de água, o

crescimento foi, em média, mais lento do que nos ratos que receberam água à vontade.

7.5.2 - Com os dados apresentados na Tabela 7.9, faça um diagrama de dispersão. Verifique que os dados não estão em torno de uma reta. Depois, faça uma transformação de variável definindo 1/X como variável explanatória. Faça um diagrama de dispersão. Verifique que os valores das variáveis 1/X e Y estão praticamente sobre uma reta. Então, ajuste uma regressão linear simples de Y contra 1/X.

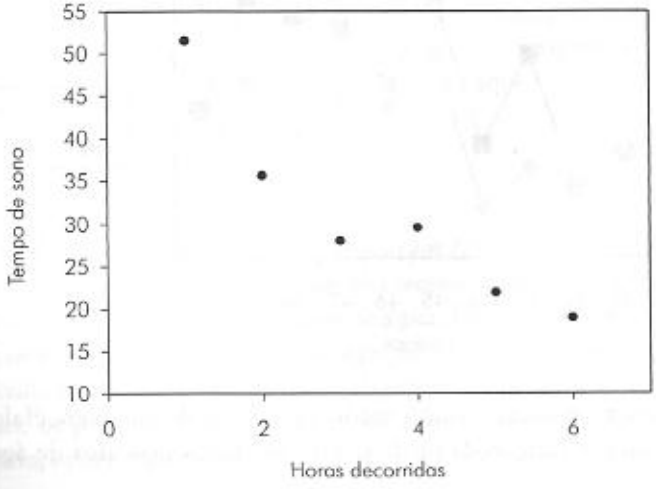
Tabela 7.9

Horas decorridas após a hepatectomia parcial (remoção de parte do fígado) e tempo de sono, em minutos, induzido por injeção intraperitoneal de 40mg de metohexital por quilo de peso vivo, em ratos

Horas decorridas X	Tempo de sono Y
24	51,57
48	35,68
120	28,05
240	29,58
480	21,87
720	18,90

Fonte: MATTOS FILHO (1976)

Figura 7.7 *Diagrama de dispersão para os dados apresentados na Tabela 7.9*



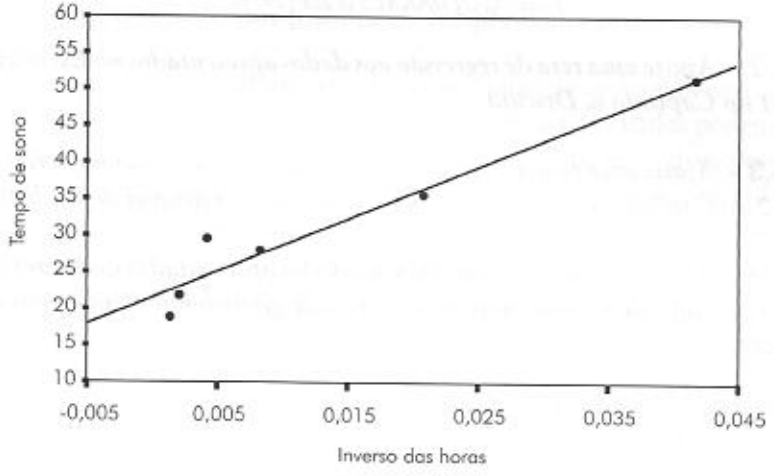
Os valores das variáveis X, $1/X$ e Y estão na Tabela 7.10, e o gráfico de Y contra $1/X$ está apresentado na Figura 7.8. É fácil ver, observando essa figura, que os pontos correspondentes aos pares de valores $1/X$ e Y estão em torno de uma reta.

Tabela 7.10

Valores de X, $1/X$ e Y, relativos aos dados apresentados na Tabela 7.9

X	$\frac{1}{X}$	Y
24	0,0417	51,57
48	0,0208	35,68
120	0,0083	28,05
240	0,0042	29,58
480	0,0021	21,87
720	0,0014	18,90

Figura 7.8 *Reta de regressão para tempo de sono dos ratos em função do inverso das horas decorridas após a hepatectomia parcial*



Para ajustar a regressão linear simples de Y contra $1/X$, é preciso calcular:

$$b = \frac{3{,}322 - \frac{0{,}0785 \cdot 185{,}65}{6}}{0{,}002264 - \frac{(0{,}0785)^2}{6}} = 722$$

$$a = 30{,}94 - 722 \cdot 0{,}01308 = 21{,}5$$

Então a reta de regressão de Y contra $1/X$ é:

$$\hat{Y} = 21{,}5 + 722\,\frac{1}{X}$$

7.6 - Exercícios Propostos

7.6.1 - *Faça um gráfico de linhas para os dados apresentados na Tabela 7.11. Discuta.*

Tabela 7.11

Idade, em dias, e peso médio, em gramas, de oito ratos machos da raça Wistar

Idade	Peso médio
30	63,94
34	74,91
38	81,65
42	95,05
46	105,89

Fonte: GUIMARÃES et alii (1979)

7.6.2 - *Ajuste uma reta de regressão aos dados apresentados no Exercício 6.5.1 do Capítulo 6. Discuta.*

7.6.3 - *Ajuste uma reta de regressão aos dados apresentados no Exercício 6.5.2 do Capítulo 6, considerando peso seco como a variável dependente.*

7.6.4 - *Ajuste uma reta de regressão aos dados apresentados no Exercício 6.5.3 do Capítulo 6, para mostrar como a idade gestacional afeta o peso ao nascer.*

CAPÍTULO
8

Noções sobre Probabilidade

Os capítulos anteriores mostram como apresentar dados e como calcular medidas que descrevem características específicas destes dados. Mas o pesquisador da área de saúde — além de fazer tabelas e gráficos, calcular médias e desvios padrões — sempre tem a pretensão de fazer *inferência*.

Para entender melhor esta afirmativa, imagine que um pesquisador anotou a idade e a pressão arterial de seus pacientes. Os dados podem ser apresentados em tabelas e gráficos, podem ser obtidas médias, desvios padrões e a reta que dá a variação da pressão arterial em função da idade. Mas este pesquisador também gostaria de estender suas conclusões a outros pacientes, além daqueles que examinou. Então este pesquisador gostaria de fazer inferência.

Para fazer inferência estatística usam-se técnicas que exigem o conhecimento de probabilidade. Neste Capítulo, e nos Capítulos 9 e 10, são dados alguns conceitos de probabilidade que preparam o leitor para entender os capítulos subseqüentes, que tratam da inferência estatística.

8.1 - Questões Básicas

Se são possíveis n eventos mutuamente exclusivos e igualmente prováveis, e se m desses eventos têm determinada característica, a *probabilidade* de que ocorra um evento com essa característica é dada pela razão m/n. O resultado pode ser multiplicado por 100, para ser dado em porcentagem.

Como exemplo, imagine que um dado será jogado. Podem ocorrer os eventos:

1, 2, 3, 4, 5 ou 6.

Esses seis eventos são mutuamente exclusivos porque duas faces não podem ocorrer ao mesmo tempo. Se o dado for honesto, os seis eventos são igualmente prováveis. Fica, então, fácil responder algumas perguntas. Por exemplo, qual é a probabilidade de sair número ímpar?

Dos seis eventos possíveis, três são ímpares. Então a probabilidade de sair número ímpar, quando se joga um dado, é:

$$\frac{3}{6} = \frac{1}{2} = 0{,}5 \text{ ou } 50\%.$$

Considere outro exemplo. Uma carta será retirada ao acaso de um baralho. Qual é a probabilidade de sair um ás? Ora, um baralho tem 52 cartas, das quais quatro são ases. Então, a probabilidade de sair um ás é:

$$\frac{4}{52} = \frac{1}{13} = 0{,}0769 \text{ ou } 7{,}69\%.$$

A probabilidade varia entre 0 e 1, ou entre 0 e 100%. Se é *certo* ocorrer determinado evento, a probabilidade desse evento é 1, ou 100%; se é *impossível* ocorrer determinado evento, a probabilidade desse evento é zero. Por exemplo, a probabilidade de ocorrer número menor do que 8, no lançamento de um dado é 1, ou 100% (evento certo). Já a probabilidade de ocorrer número maior do que 8 é zero (evento impossível).

8.2 - Probabilidade Condicional

A idéia de probabilidade condicional pode ser entendida através de um exemplo. Imagine que um dado foi jogado. Qual é a probabilidade de ter ocorrido 5? Como o dado tem seis faces, a probabilidade de ter ocorrido a face com número 5 é

$$\frac{1}{6} = 0{,}1667 \text{ ou } 16{,}67\%.$$

Imagine agora que o dado foi jogado e já se sabe que ocorreu face com número ímpar. Qual é a probabilidade de ter ocorrido 5? Note que a resposta a esta pergunta é diferente da resposta dada à pergunta anterior. Se saiu face com número ímpar, só podem ter ocorrido os números: 1, 3 ou 5. Logo, a probabilidade de ter ocorrido 5 é:

$$\frac{1}{3} = 0{,}3333 \text{ ou } 33{,}33\%.$$

A probabilidade de ocorrer determinado evento pode ser *modificada* quando se impõe uma condição. Como mostra o exemplo, a probabilidade de ocorrer 5 no jogo de um dado é 16,67%, mas, *sob a condição* de ter ocorrido face com número ímpar, a probabilidade de ocorrer 5 é 33,33%.

Denomina-se *probabilidade condicional* à probabilidade de ocorrer determinado evento sob uma dada condição. Indica-se a probabilidade condicional de ocorrer o evento A sob a condição de ter ocorrido B por P(A|B), que se lê "probabilidade de A dado B".

Como exemplo, considere a probabilidade de ocorrer um acidente automobilístico, dado que está chovendo. Esta probabilidade é *condicional*, porque se refere à probabilidade de ocorrer um evento (acidente) *sob uma dada condição* (estar chovendo).

8.3 - Eventos Independentes

Para entender a idéia de eventos independentes, imagine que um dado e uma moeda são jogados ao mesmo tempo e se pergunte: a) qual é a probabilidade de ocorrer cara na moeda? b) qual é a probabilidade de ocorrer cara na moeda sabendo que ocorreu face 6 no dado?

Na Tabela 8.1 estão os eventos que podem ocorrer quando se jogam um dado e uma moeda ao mesmo tempo.

Tabela 8.1

Eventos possíveis no jogo de um dado e uma moeda

Dado	Moeda	
	Cara	Coroa
1	Cara; 1	Coroa; 1
2	Cara; 2	Coroa; 2
3	Cara; 3	Coroa; 3
4	Cara; 4	Coroa; 4
5	Cara; 5	Coroa; 5
6	Cara; 6	Coroa; 6

Dos 12 eventos possíveis e igualmente prováveis apresentados na Tabela 8.1, seis correspondem à saída de cara na moeda. Então a probabilidade de sair cara na moeda é:

$$\frac{6}{12} = \frac{1}{2} = 0{,}5 \text{ ou } 50\%.$$

Para obter a probabilidade de sair cara na moeda, sabendo que saiu 6 no dado, observe a última linha da Tabela 8.1. Dos dois eventos que correspondem à saída de 6 no dado, um corresponde à saída de cara na moeda. Então a probabilidade de sair cara na moeda, sabendo que ocorreu 6 no dado, é:

$$\frac{1}{2} = 0{,}5 \text{ ou } 50\%.$$

Neste exemplo, a probabilidade de ocorrer um evento (sair cara na moeda) não foi modificada pela ocorrência de outro evento (sair 6 no dado). Diz-se então que esses eventos são independentes.

Por definição, dois eventos são *independentes* quando a probabilidade de ocorrer um deles não é modificada pela ocorrência do outro. Quando se jogam um dado e uma moeda, o resultado que ocorre na moeda não depende do que ocorre no dado. Então esses eventos são independentes. Escreve-se $P(A \mid B) = P(A)$.

Na área biológica existem vários exemplos de "eventos dependentes" e de "eventos independentes". Assim, "olhos claros" e "cabelos claros" são eventos dependentes porque a probabilidade de uma pessoa ter olhos claros é maior se a pessoa tem cabelos claros. Já "olhos claros" e "idade avançada" são eventos independentes, porque a probabilidade de uma pessoa ter olhos claros não aumenta (ou diminui) com a idade.

8.4 - Teorema do Produto

Uma moeda será jogada duas vezes. Qual é a probabilidade de ocorrer cara nas duas jogadas? Ora, a probabilidade de ocorrer cara na primeira jogada é:

$$\frac{1}{2} = 0{,}5 \text{ ou } 50\%.$$

A probabilidade de ocorrer cara na segunda jogada também é:

$$\frac{1}{2} = 0{,}5 \text{ ou } 50\%,$$

porque o fato de ocorrer cara na primeira jogada não modifica a probabilidade de ocorrer cara na segunda jogada (os eventos são independentes).

Para obter a probabilidade de ocorrer cara nas duas jogadas (primeira e segunda), faz-se o produto:

$$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = 0{,}25 \text{ ou } 25\%.$$

Veja agora outro problema: uma urna contém duas bolas brancas e uma vermelha. Retiram-se duas bolas da urna ao acaso, uma em seguida da outra e sem que a primeira tenha sido recolocada. Qual é a probabilidade de as duas serem brancas?

A probabilidade de a primeira bola ser branca é:

$$\frac{2}{3} = 0{,}6667 \text{ ou } 66{,}67\%.$$

A probabilidade de a segunda bola ser branca depende do que ocorreu na primeira retirada. Se saiu bola branca, a probabilidade de a segunda também ser branca é:

$$\frac{1}{2} = 0{,}5 \text{ ou } 50\%.$$

Para obter a probabilidade de as duas bolas retiradas serem brancas, faz-se o produto:

$$\frac{2}{3} \cdot \frac{1}{2} = \frac{2}{6} = \frac{1}{3} = 0{,}3333 \text{ ou } 33{,}33\%.$$

Agora fica fácil entender o teorema do produto. Se A e B são eventos independentes, a probabilidade de ocorrer A e B é dada pela probabilidade de ocorrer A, multiplicada pela probabilidade de ocorrer B. Escreve-se:

$$P(A \text{ e } B) = P(A) \cdot P(B).$$

Se A e B não são independentes, a probabilidade de ocorrer A e B é dada pela probabilidade de ocorrer A, multiplicada pela probabilidade (condicional) de ocorrer B, dado que A ocorreu. Escreve-se:

$$P(A \text{ e } B) = P(A) \cdot P(B|A).$$

8.5 - Teorema da Soma

Fica mais fácil entender o teorema da soma com a ajuda de exemplos. Suponha então que uma urna contém duas bolas brancas, uma azul e uma vermelha. Retira-se uma bola da urna ao acaso. Qual a probabilidade de ter saído bola colorida, isto é, azul ou vermelha? Ora, a probabilidade de sair bola azul é:

$$\frac{1}{4} = 0{,}25 \text{ ou } 25\%.$$

e a probabilidade de sair bola vermelha é:

$$\frac{1}{4} = 0{,}25 \text{ ou } 25\%.$$

Então a probabilidade de sair bola colorida, isto é, azul ou vermelha, é dada pela soma:

$$\frac{1}{4}+\frac{1}{4} = \frac{2}{4} = \frac{1}{2} = 0{,}5 \text{ ou } 50\%.$$

Imagine agora que uma carta será retirada ao acaso de um baralho. Qual é a probabilidade de sair uma carta de espadas ou um ás?

Como um baralho tem 52 cartas, das quais 13 são de espadas e quatro são ases, alguém poderia pensar que a probabilidade de sair uma carta de espadas ou um ás é dada pela soma:

$$\frac{13}{52}+\frac{4}{52},$$

mas esta resposta estaria errada, porque existe uma carta, o ás de espadas, que é tanto ás como espadas. Então o ás de espadas teria sido contado duas vezes. A probabilidade de sair uma carta de espadas ou um ás é dada por

$$\frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13} = 0{,}3077 \text{ ou } 30{,}77\%.$$

Agora fica fácil entender o teorema da soma. Se os eventos A e B não podem ocorrer ao mesmo tempo, a probabilidade de ocorrer A ou B é dada pela probabilidade de A, mais a probabilidade de B. Escreve-se:

$$P(A \text{ ou } B) = P(A) + P(B).$$

Se A e B podem ocorrer ao mesmo tempo, a probabilidade de ocorrer A ou B é dada pela probabilidade de A, mais a probabilidade de B, menos a probabilidade de A e B. Escreve-se:

$$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ e } B).$$

8.6 - Exercícios Resolvidos

8.6.1 - *Um casal tem dois filhos. Qual é a probabilidade de: a) o primogênito ser homem? b) os dois filhos serem homens? c) pelo menos um dos filhos ser homem?*

Suponha que a probabilidade de nascer menino é $\frac{1}{2}$ e que o sexo do segundo filho não depende do sexo do primeiro. Então:

a) a probabilidade de o primogênito ser homem é:

$$\frac{1}{2} \text{ ou } 50\%;$$

b) a probabilidade de os dois filhos serem homens é:

$$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \text{ ou } 25\%;$$

c) a probabilidade de pelo menos um dos filhos ser homem é

$$\frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4} \text{ ou } 75\%,$$

porque são quatro eventos possíveis: (menino-menino), (menino-menina), (menina-menino), (menina-menina), dos quais os três primeiros atendem à característica "pelo menos um dos filhos ser homem".

8.6.2 - *No cruzamento de ervilhas amarelas homozigotas (AA) com ervilhas verdes homozigotas (aa) ocorrem ervilhas amarelas heterozigotas (Aa). Se estas ervilhas forem cruzadas entre si, ocorrem ervilhas amarelas e verdes, na proporção de três para uma. Suponha que foram pegas, ao acaso, três ervilhas resultantes do cruzamento de ervilhas amarelas heterozigotas. Qual a probabilidade de as três serem verdes?*

A probabilidade de uma ervilha resultante do cruzamento Aa x Aa ser verde é 1/4. Logo, a probabilidade de as três serem verdes é:

$$\frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{4^3} = \frac{1}{64} \text{ ou } 1{,}56\%.$$

8.7 - Exercícios Propostos

8.7.1 - *Um casal tem dois filhos. Qual é a probabilidade de: a) o segundo filho ser homem? b) o segundo filho ser homem, dado que o primeiro é homem?*

8.7.2 - *A probabilidade de determinado teste para a AIDS dar resultado negativo em portadores de anticorpos contra o vírus (falso negativo) é 10%. Supondo que falsos negativos ocorrem independentemente, qual é a probabilidade de um portador de anticorpos contra o vírus da AIDS,*

que se apresentou três vezes para o teste, ter tido, nas três vezes, resultado negativo?

8.7.3 - *Uma pessoa normal, filha de pais normais, tem um avô albino (aa). Se os outros avós não forem portadores do gene para albinismo (AA), qual é a probabilidade de essa pessoa ser portadora do gene para albinismo (Aa)?*

8.7.4 - *Suponha que a probabilidade de uma pessoa ser do tipo sangüíneo O é 40%, ser A é 30% e ser B é 20%. Suponha ainda que a probabilidade de Rh^+ é de 90% e que o fator Rh independe do tipo sangüíneo. Nestas condições, qual é a probabilidade de uma pessoa tomada ao acaso da população ser: a) O, Rh^+? b) AB, Rh^-?*

8.7.5 - *Qual é a probabilidade de ser hemofílico (X_bY) o filho de um homem normal (XY) e de uma filha de hemofílico (X_bX)?*

CAPÍTULO
9

Distribuição Binomial

A distribuição binomial apresentada neste capítulo será usada para explicar os testes estatísticos. Mas antes de explicar a distribuição binomial, é preciso conceituar variável aleatória e distribuição discreta.

9.1 - Variável Aleatória

Imagine que um laboratório cria ratos de uma só raça e mesma progênie, em condições controladas de alimentação e manejo. É razoável considerar que os pesos desses ratos variam. Sabe-se que os machos pesam mais do que as fêmeas e que os animais ganham peso com a idade. No entanto, mesmo ratos de um só sexo, nascidos no mesmo dia, têm pesos variáveis.

Essa variabilidade ocorre ao acaso, pois resulta de uma soma de fatores não-controlados. Toda vez que uma variável é influenciada pelo acaso, diz-se que é uma *variável aleatória*. No exemplo, o peso dos ratos de mesmo sexo e mesma idade varia ao acaso. Então, peso de ratos é uma variável aleatória.

As variáveis aleatórias são indicadas por letras maiúsculas. Então o peso dos ratos pode ser indicado pela letra X. Os valores assumidos pelas variáveis aleatórias são indicados por letras minúsculas. Então o valor que se obtém quando se pesa determinado rato é indicado pela letra x.

As variáveis aleatórias podem ser discretas ou contínuas. A *variável aleatória é discreta*, quando só assume valores que podem ser associados aos números naturais (1, 2, 3, 4 etc.). Por exemplo, o número de pacientes atendidos por dia, em uma clínica, é uma variável aleatória discreta.

Existe um tipo especial de variável aleatória discreta: é aquela que só assume um de dois valores possíveis. Este tipo de variável recebe, em estatística, o nome de *variável aleatória binária*. O exemplo clássico de variável aleatória binária é o resultado que ocorre no jogo de uma moeda, que só pode ser cara ou coroa. Associa-se o número zero a um dos valores da variável aleatória binária e o número 1 ao outro valor. Por exemplo, quando se joga uma moeda pode-se convencionar que coroa vale zero e cara vale 1.

Na área de saúde ocorrem muitas variáveis aleatórias binárias. Por exemplo, o médico pode classificar um paciente como chagásico ou não-chagásico, um recém-nascido como portador ou não-portador de anomalia congênita, uma amostra de sangue como do tipo Rh^+ ou Rh^-.

As *variáveis aleatórias contínuas* assumem infinitos valores em um dado intervalo. Por exemplo, peso corporal é uma variável aleatória contínua porque, em princípio, pode assumir infinitos valores em um dado intervalo. Na prática, porém, o número de valores de peso que podem ser distinguidos em um dado intervalo é limitado pela precisão da balança.

9.2 - Distribuição Discreta

Entende-se por *distribuição discreta* o conjunto de todos os valores que podem ser assumidos pela variável aleatória discreta, com as respectivas probabilidades. Então os resultados que podem ocorrer no jogo de um dado, com as respectivas probabilidades, constituem um exemplo de distribuição discreta. Veja a Tabela 9.1.

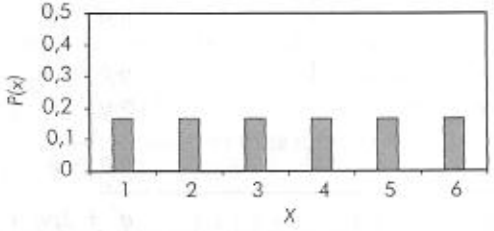
Tabela 9.1
Distribuição dos resultados do jogo de um dado

X	$P(X)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
Total	1

A distribuição discreta pode ser apresentada graficamente. Para isso, primeiro se traça o sistema de eixos cartesianos. Depois se colocam,

no eixo das abscissas, todos os valores possíveis da variável aleatória e, no eixo das ordenadas, as respectivas probabilidades. Veja, como exemplo, a Figura 9.1, que apresenta graficamente a distribuição dada na Tabela 9.1.

Figura 9.1 *Distribuição dos resultados do jogo de um dado*



A soma das probabilidades associadas a todos os valores possíveis de uma variável aleatória é sempre igual a 1. Note que isto acontece no exemplo apresentado na Tabela 9.1.

9.3 - Distribuição Binomial

A *distribuição binomial* é uma distribuição discreta que resulta da soma de variáveis aleatórias binárias. Para estudar a distribuição binomial, será dado um exemplo. Considere então a variável que representa o sexo de um nascituro. Essa variável é aleatória binária porque um nascituro só pode ser menino ou menina. Considere ainda que, se o nascituro for menino, a variável assume valor 1 e, se for menina, assume valor zero.

Para estudar a distribuição do número de meninos em n nascimentos, é preciso fazer $n = 1, 2, 3$ etc. Seja $p = \frac{1}{2}$ a probabilidade de um nascituro ser menino, e $q =$ a probabilidade de um nascituro ser menina. É claro que $p + q = 1$. Para $n = 1$, isto é, no caso de um único nascimento, o número de meninos só pode ser zero ou 1, como mostra a Tabela 9.2.

Tabela 9.2
Distribuição do número de meninos em um nascimento

Número de nascimentos	Número de meninos	Probabilidade
1	0	q
	1	p
Total		$q+p = 1$

Para $n = 2$, isto é, no caso de dois nascimentos, o número de meninos pode ser zero, 1 ou 2, como mostra a Tabela 9.3.

Tabela 9.3

Distribuição do número de meninos em 2 nascimentos

Número de nascimentos	Número de meninos	Probabilidade
2	0	$q \cdot q = q^2$
	1	$\left.\begin{matrix} q \cdot p \\ p \cdot q \end{matrix}\right\} = 2pq$
	2	$p \cdot p = p^2$
Total		$q^2 + 2pq + p^2 = 1$

Para $n = 3$, isto é, no caso de três nascimentos, o número de meninos pode ser zero, 1, 2 ou 3, como mostra a Tabela 9.4.

Tabela 9.4

Distribuição do número de meninos em 3 nascimentos

Número de nascimentos	Número de meninos	Probabilidade
3	0	$q \cdot q \cdot q = q^3$
	1	$\left.\begin{matrix} q \cdot q \cdot p \\ q \cdot p \cdot q \\ p \cdot q \cdot q \end{matrix}\right\} = 3q^2p$
	2	$\left.\begin{matrix} q \cdot p \cdot p \\ p \cdot q \cdot p \\ p \cdot p \cdot q \end{matrix}\right\} = 3qp^2$
	3	$p \cdot p \cdot p = p^3$
Total		$q^3 + 3p^2q + 3pq^2 + p^3 = 1,$

O número de meninos, em n nascimentos, é uma soma de variáveis aleatórias binárias. Essa soma tem *distribuição binomial*. Por exemplo, para $n = 2$, o número de meninos pode ser zero (0 + 0), 1 (0 + 1 ou 1 + 0) ou 2 (1 + 1).

A distribuição binomial fica definida quando são dados dois parâmetros: o número (n) de variáveis aleatórias binárias observadas (por exemplo, 2 nascimentos) e a probabilidade (p) de ocorrer valor 1 em uma única observação (no exemplo, de nascer menino, quando se considera um único nascimento). Dados n e p, a probabilidade de a variável aleatória assumir valor x é obtida pela expressão:

$$P(x) = \binom{n}{x} p^x q^{n-x}$$

onde $\binom{n}{x}$ é a combinação de n, x a x. Uma rápida revisão sobre análise combinatória é dada no Capítulo 15 deste livro.

Para entender a aplicação da fórmula, imagine que se deseja obter a probabilidade de ocorrerem 2 meninos em 4 nascimentos. Como $n = 4$ nascimentos, a probabilidade de ocorrer menino em um nascimento é $p = 1/2$ e a probabilidade de ocorrer uma menina é $q = 1/2$, segue-se que a probabilidade de x ser igual a 2 é:

$$P(2) = \binom{4}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2$$

$$= \frac{4!}{2!(4-2)!} \cdot \frac{1}{2^2} \cdot \frac{1}{2^2}$$

$$= 6 \cdot \frac{1}{4} \cdot \frac{1}{4}$$

$$= 0{,}375 \text{ ou } 37{,}5\%$$

Considere outro exemplo. O resultado do cruzamento de ervilhas amarelas homozigotas (AA) com ervilhas verdes homozigotas (aa) são ervilhas amarelas heterozigotas (Aa). Se estas ervilhas forem cruzadas entre si, ocorrem ervilhas amarelas e verdes, na proporção de 3 para 1. Portanto, a probabilidade de, num cruzamento desse tipo, ocorrer ervilha amarela é $p = 3/4$ e a probabilidade de ocorrer ervilha verde é $q = 1/4$.

Com base no que foi visto até aqui, pode-se entender que o número de ervilhas amarelas em um conjunto de n ervilhas é uma variável aleatória com distribuição binomial de parâmetros n e $p = 3/4$. Imagine que foram pegas, ao acaso, 4 ervilhas resultantes do cruzamento de ervilhas amarelas heterozigotas. Qual é a probabilidade de 2 dessas 4 ervilhas serem de cor amarela?

Ora, a probabilidade de uma ervilha ser amarela é $p = 3/4$ e de ser verde é $q = 1/4$. Então, a probabilidade de 2 das 4 ervilhas serem amarelas é dada por:

$$P(2) = \binom{4}{2} \left(\frac{3}{4}\right)^2 \left(\frac{1}{4}\right)^2$$

$$= \frac{4!}{2!(4-2)!} \cdot \frac{3^2}{4^2} \cdot \frac{1}{4^2}$$

$$= 6 \cdot \frac{9}{16} \cdot \frac{1}{16}$$

$$= 0{,}2109 \text{ ou } 21{,}09\%$$

9.4 - Média e Variância na Distribuição Binomial

A média μ (lê-se mi) de uma distribuição binomial é dada pela fórmula:

$$\mu = np$$

e a variância σ^2 (lê-se sigma ao quadrado) é dada pela fórmula:

$$\sigma^2 = npq$$

É fácil calcular a média e a variância de uma distribuição binomial. Como exemplo, considere o número de meninos em 1 000 nascituros. Essa variável tem distribuição binomial. Então, em 1 000 nascituros ocorrem, em média:

$$\mu = np = 1\,000 \cdot \frac{1}{2} = 500 \text{ meninos,}$$

e a variância é

$$\sigma^2 = npq = 1\,000 \cdot \frac{1}{2} \cdot \frac{1}{2} = 250.$$

Como outro exemplo, considere o cruzamento de ervilhas amarelas e verdes, descrito anteriormente. Um conjunto de $n = 100$ ervilhas tem, em média:

$$\mu = 100 \cdot \frac{3}{4} = 75 \text{ ervilhas amarelas}$$

e a variância é:

$$\sigma^2 = 100 \cdot \frac{3}{4} \cdot \frac{1}{4} = 18{,}75.$$

9.5 - Exercícios Resolvidos

9.5.1 - ***A probabilidade de um menino ser daltônico é 8%. Qual é a probabilidade de serem daltônicos todos os 4 meninos que se apresentaram, em determinado dia, para um exame oftalmológico?***

No problema, $p = 0{,}08$. Então $q = 1 - 0{,}08 = 0{,}92$. O número de meninos é $n = 4$. Para obter a probabilidade de x assumir valor 4, aplica-se a fórmula:

$$P(x) = \binom{n}{x} p^x q^{n-x}$$

Então:

$$P(4) = \binom{4}{4} (0{,}08)^4 (0{,}92)^0$$

$$= \frac{4!}{4!(4-4)!} \cdot (0{,}08)^4 (0{,}92)^0$$

$$= 0{,}00004096 \text{ ou } 0{,}004096\%$$

9.5.2 - *Apresente, em tabela e em gráfico, a distribuição do número de meninos que podem ocorrer em uma família com 6 crianças.*

No problema, n é o número de crianças (6), p é a probabilidade de menino (1/2) e q é a probabilidade de menina (1/2). Para obter a probabilidade de X assumir o valor 0, ou seja, de não ocorrer nenhum menino, calcule:

$$P(0) = \binom{6}{0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^6$$

$$= \frac{6!}{0!(6-0)!} \cdot \frac{1}{2^0} \cdot \frac{1}{2^6} = \frac{1}{2^6} = \frac{1}{64}$$

Para obter a probabilidade de X assumir o valor 1, isto é, de ocorrer um menino em uma família com 6 crianças, calcule:

$$P(1) = \binom{6}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^5$$

$$= \frac{6!}{1!(6-1)!} \cdot \frac{1}{2^1} \cdot \frac{1}{2^5} = 6 \cdot \frac{1}{2^6} = \frac{6}{64}$$

Para obter a probabilidade de x assumir o valor 2, isto é, de ocorrerem dois meninos em uma família com 6 crianças, calcule:

$$P(2) = \binom{6}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^4$$

$$= \frac{6!}{2!(6-2)!} \cdot \frac{1}{2^2} \cdot \frac{1}{2^4} = \frac{6 \cdot 5}{2} \cdot \frac{1}{2^6} = \frac{15}{64}$$

Para obter a probabilidade de X assumir o valor 3, calcule:

$$P(3) = \binom{6}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^3$$

$$= \frac{6!}{3!(6-3)!}\cdot\frac{1}{2^3}\cdot\frac{1}{2^3} = \frac{6\cdot5\cdot4}{2\cdot3}\cdot\frac{1}{2^6} = \frac{20}{64}$$

Para obter a probabilidade de X assumir o valor 4, calcule:

$$P(4) = \binom{6}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^2$$

$$= \frac{6!}{4!\,(6-4)!}\cdot\frac{1}{2^4}\cdot\frac{1}{2^2} = \frac{6\cdot5}{2}\cdot\frac{1}{2^6} = \frac{15}{64}$$

Para obter a probabilidade de X assumir o valor 5, calcule:

$$P(5) = \binom{6}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^1$$

$$= \frac{6!}{5!(6-5)!}\cdot\frac{1}{2^5}\cdot\frac{1}{2} = 6\cdot\frac{1}{2^6} = \frac{6}{64}$$

Para obter a probabilidade de X assumir o valor 6, calcule:

$$P(6) = \binom{6}{6}\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^0$$

$$= \frac{6!}{6!\,(6-6)!}\cdot\frac{1}{2^6} = \frac{1}{64}$$

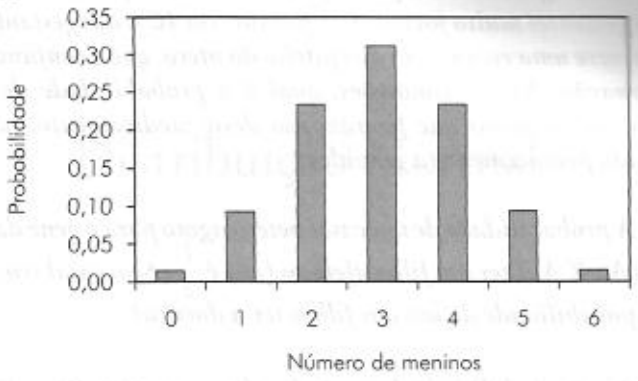
Calculadas as probabilidades associadas a todos os valores possíveis de X, organiza-se a distribuição apresentada na Tabela 9.5. Com os valores apresentados na Tabela 9.5 constrói-se o gráfico de barras da Figura 9.2.

Tabela 9.5

Distribuição do número de meninos em uma família com 6 crianças

Evento	X	$P(X)$
Nenhum menino	0	1/64
1 menino	1	6/64
2 meninos	2	15/64
3 meninos	3	20/64
4 meninos	4	15/64
5 meninos	5	6/64
6 meninos	6	1/64

Figura 9.2 *Distribuição do número de meninos em uma família com 6 crianças*



9.5.3 - Um exame é constituído de 100 testes com 5 alternativas, onde apenas uma é correta. Quantos testes acerta, em média, um aluno que nada sabe sobre a matéria do exame? Qual é a variância da distribuição?

A probabilidade de um aluno acertar uma resposta por acaso é $p = 1/5$. Existem $n = 100$ testes. Então, aplicando a fórmula, vem:

$$\mu = np = 100 \cdot \frac{1}{5} = 20$$

ou seja, um aluno que nada sabe sobre a matéria acerta em média 20 testes. A variância da distribuição é:

$$\sigma^2 = npq = 100 \cdot \frac{1}{5} \cdot \frac{4}{5} = 16.$$

9.6 - Exercícios Propostos

9.6.1 - Seja X a variável aleatória que indica o número de meninos em uma família com 5 crianças. Apresente a distribuição de X em uma tabela. Faça um gráfico.

9.6.2 - Um exame é constituído de dez testes tipo certo-errado. Quantos testes acerta, em média, um aluno que nada sabe sobre a matéria do exame? Qual é a variância da distribuição?

9.6.3 - Um exame é constituído de dez testes com 5 alternativas, onde apenas uma é correta. Quantos testes acerta, em média, um aluno que nada sabe sobre a matéria do exame? Qual é a variância da distribuição?

9.6.4 - *Suponha que determinado medicamento usado para o diagnóstico precoce da gravidez é capaz de confirmar casos positivos em apenas 90% das gestantes muito jovens. Isto porque, em 10% das gestantes muito jovens, ocorre uma escamação do epitélio do útero, que é confundida com a menstruação. Nestas condições, qual é a probabilidade de 2, de 3 gestantes muito jovens que fizeram uso desse medicamento, não terem confirmado precocemente a gravidez?*

9.6.5 - *A probabilidade de um casal heterozigoto para o gene da fenilcetonúria (Aa X Aa) ter um filho afetado (aa) é* $\frac{1}{4}$ *. Se o casal tem 3 filhos, qual é a possibilidade de um dos filhos ter a doença?*

9.6.6 - *Se a probabilidade de um indivíduo ter sangue* Rh^- *é 10%, qual é a possibilidade de 5 indivíduos que se apresentaram para exame de sangue serem todos* Rh^-*?*

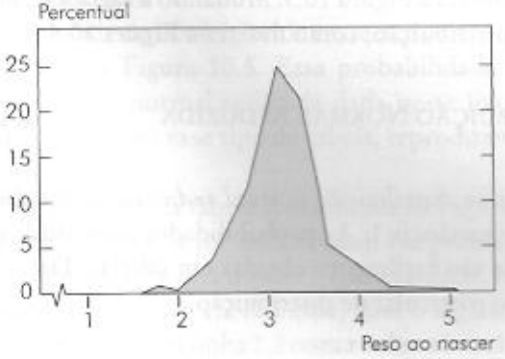
CAPÍTULO
10

Distribuição Normal

O pesquisador estuda variáveis. O estatístico diz que essas variáveis são aleatórias porque elas têm um componente que varia ao acaso. Por exemplo, a variabilidade dos pesos ao nascer de nascidos vivos de mesmo sexo, mesma raça, mesma idade gestacional e filhos de mães em condições similares de saúde e alimentação é explicada pelo acaso. Então o peso ao nascer é uma variável aleatória.

As grandes amostras de certas variáveis aleatórias permitem construir gráficos que têm aparência típica. Como exemplo, observe a Figura 10.1, que apresenta uma distribuição de pesos ao nascer de nascidos vivos brancos de sexo masculino, com cerca de 40 semanas de gestação. Gráficos com esse tipo de configuração são obtidos, por exemplo, quando se analisa o peso ao nascer de cerca de 2 000 nascidos com iguais características.

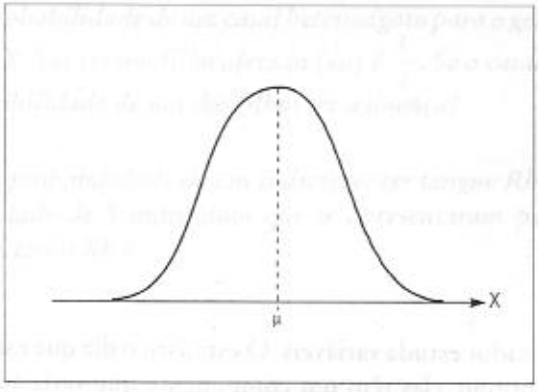
Figura 10.1 *Peso ao nascer de nascidos vivos brancos do sexo masculino com cerca de 40 semanas de gestação*



10.1 - Características Gerais

As medidas biológicas, as medidas de produtos fabricados em série e os erros de medidas dão origem a gráficos semelhantes ao apresentado na Figura 10.1. Todas essas medidas são variáveis que têm distribuições que se aproximam da *distribuição normal*, apresentada na Figura 10.2.

Figura 10.2 *Gráfico da distribuição normal*



A distribuição normal tem as seguintes características:

a) a variável aleatória pode assumir qualquer valor real;
b) o gráfico da distribuição normal é uma curva em forma de sino, simétrica em torno da média μ (lê-se mi), como mostra a Figura 10.2;
c) a área total sob a curva vale 1, porque essa área corresponde à probabilidade de a variável aleatória assumir qualquer valor real;
d) como a curva é simétrica em torno da média, os valores maiores do que a média e os valores menores do que a média ocorrem com igual probabilidade;
e) a configuração da curva é dada por dois *parâmetros*: a média μ e a variância σ^2. Mudando a média, muda a posição da distribuição, como mostra a Figura 10.3. Mudando a variância, muda a dispersão da distribuição, como mostra a Figura 10.4.

10.2 - Distribuição Normal Reduzida

Denomina-se *distribuição normal reduzida* a distribuição normal de média zero e variância 1. As probabilidades associadas à distribuição normal reduzida são facilmente obtidas em tabelas. Daí o interesse em estudar esse tipo particular de distribuição.

Figura 10.3 *Duas distribuições normais de mesma variância e com médias diferentes*

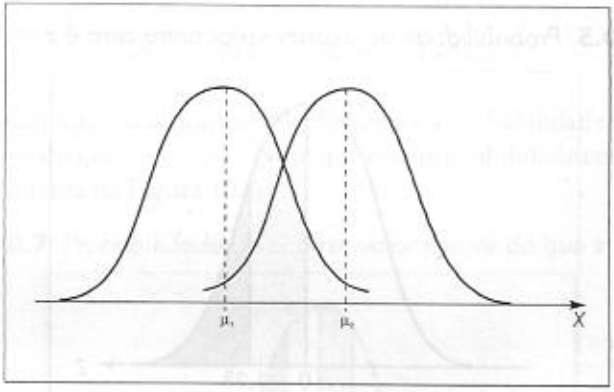
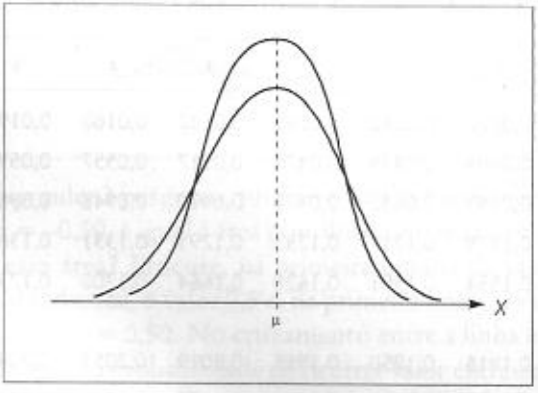


Figura 10.4 *Duas distribuições normais de mesma média e com variâncias diferentes*



Observe a Figura 10.5. A área total sob a curva vale 1. Isto significa que a probabilidade de ocorrer qualquer valor real é 1. A curva é simétrica em torno da média zero. Então a probabilidade de ocorrer valor maior do que zero é 0,5. Mas qual seria a probabilidade de ocorrer valor entre zero e $z = 1{,}25$, por exemplo?

A probabilidade de ocorrer valor entre zero e $z = 1{,}25$ corresponde à área pontilhada na Figura 10.5. Essa probabilidade é encontrada na tabela de distribuição normal reduzida dada neste livro, em Apêndice. Para mostrar como se usa esse tipo de tabela, reproduziu-se parte dela na Figura 10.6.

Na primeira *coluna* da tabela apresentada na Figura 10.6 está o valor 1,2 (para facilitar, este valor foi sombreado). Na primeira *linha* da tabela apresentada na Figura 10.6 está o valor 5 (para facilitar, este valor também foi sombreado). O número 1,2 compõe, com o algarismo 5, o número $z = 1{,}25$. No cruzamento da linha 1,2 com a coluna 5 está o número 0,3944

(também sombreado). Esta é a probabilidade de ocorrer valor entre zero e $z = 1{,}25$, que corresponde à área pontilhada na Figura 10.5.

Figura 10.5 *Probabilidade de ocorrer valor entre zero e z = 1,25*

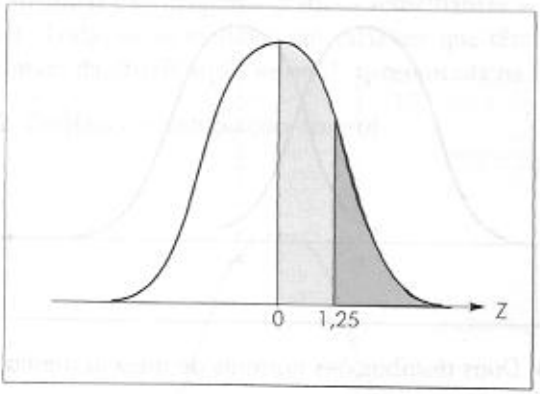


Figura 10.6 *Probabilidade de ocorrer valor entre zero e 1,25*

	0	1	2	3	4	5	6
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454
0,7	0,2580	0,2611	0,2642	0,2673	0,2703	0,2734	0,2764
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279

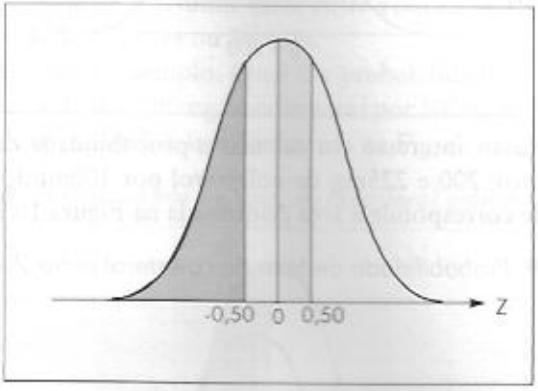
Considere outro problema. Qual é a probabilidade de ocorrer valor maior do que $z = 1{,}25$? Essa probabilidade corresponde à área hachurada na Figura 10.5. Como a probabilidade de ocorrer valor maior do que zero

é 0,5 e a probabilidade de ocorrer valor entre zero e $z = 1,25$ (área pontilhada) é 0,3944, a probabilidade pedida (área hachurada) é:

$$0,5 - 0,3944 = 0,1056 \text{ ou } 10,56\%$$

Finalmente, um último problema. Qual é a probabilidade de ocorrer valor menor do que $z = -0,50$? Note que essa probabilidade corresponde à área hachurada na Figura 10.7.

Figura 10.7 *Probabilidade de ocorrer valor menor do que $z = -0,50$*



Para responder à pergunta, primeiro observe que a área em branco, entre zero e $z = -0,50$, é igual à área pontilhada entre zero e $z = 0,50$. Mas quanto vale essa área? Procure, na primeira coluna da tabela de distribuição normal reduzida, o valor 0,5 e, na primeira linha, o valor zero, para compor o número $z = 0,50$. No cruzamento entre a linha e a coluna está o valor 0,1915, que é a probabilidade de ocorrer valor entre zero e $z = 0,50$.

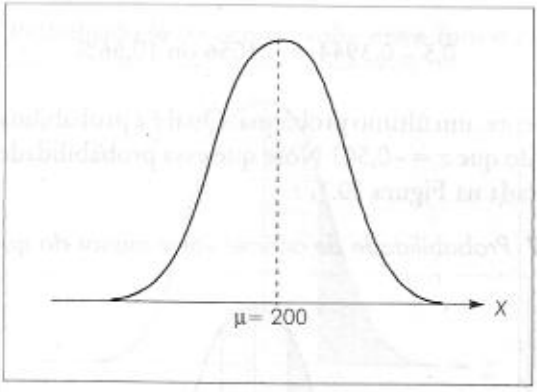
Observe novamente a Figura 10.7. A probabilidade de ocorrer valor menor do que $z = -0,50$ é igual à probabilidade de ocorrer valor maior do que $z = 0,50$. Como a probabilidade de ocorrer valor maior do que a média zero é 0,5, a probabilidade pedida é dada por:

$$0,5 - 0,1915 = 0,3085 \text{ ou } 30,85\%.$$

10.3 - Probabilidades na distribuição Normal

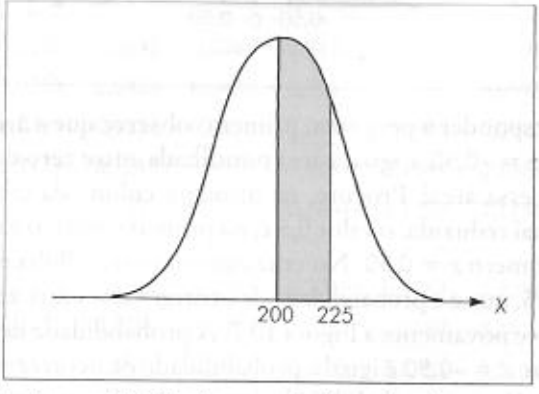
Suponha que a quantidade de colesterol em 100ml de plasma sangüíneo humano tem distribuição normal com média 200mg e desvio padrão 20mg. O gráfico dessa distribuição está apresentado na Figura 10.8.

Figura 10.8 *Distribuição normal da taxa de colesterol no plasma sangüíneo humano*



Pode existir interesse em calcular a probabilidade de uma pessoa apresentar entre 200 e 225mg de colesterol por 100ml de plasma. Essa probabilidade corresponde à área pontilhada na Figura 10.9.

Figura 10.9 *Probabilidade de taxa de colesterol entre 200 e 225*



Para calcular probabilidades associadas à distribuição normal, usa-se um artifício. Sabe-se que, se X tem distribuição normal com a média μ e desvio padrão σ, a variável

$$Z = \frac{X - \mu}{\sigma}$$

tem *distribuição normal reduzida*. Fica então fácil obter as probabilidades associadas a qualquer distribuição normal: basta "reduzir" a distribuição e obter as probabilidades na tabela de distribuição normal reduzida, como mostra a Seção 10.2.

No exemplo em discussão, deseja-se obter a probabilidade de uma pessoa apresentar entre 200 e 225mg de colesterol por 100ml de plasma. Como a quantidade de colesterol tem distribuição normal com média $\mu = 200$ e desvio padrão $\sigma = 20$, a variável

$$Z = \frac{X - 200}{20}$$

tem distribuição normal reduzida.

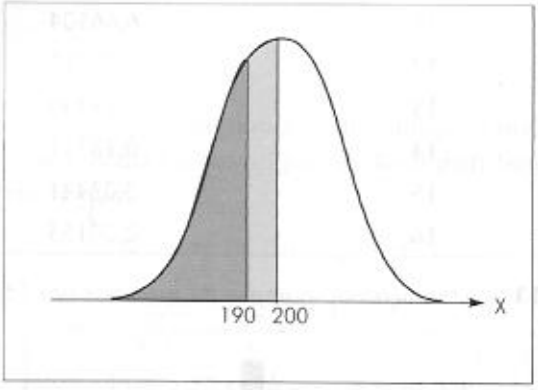
Nessa distribuição a média é zero e, ao valor $x = 225$, corresponde

$$z = \frac{225 - 200}{20} = 1{,}25$$

A área pontilhada na Figura 10.9 corresponde à área pontilhada na Figura 10.5. Então a probabilidade de X assumir valor entre 200 e 225 é igual à probabilidade de Z assumir valor entre zero e $z = 1{,}25$ que, como se viu na Seção 10.2, é 0,3944 ou 39,44%.

Considere outro exemplo. Qual é a probabilidade de uma pessoa apresentar menos do que 190mg de colesterol por 100ml de plasma? Essa probabilidade corresponde à área hachurada na Figura 10.10.

Figura 10.10 *Probabilidade de taxa de colesterol menor do que 190*



Para resolver o problema, é preciso "reduzir" o valor $x = 190$. Obtém-se então

$$z = \frac{190 - 200}{20} = -0{,}50$$

A probabilidade pedida corresponde à probabilidade de Z assumir valor menor do que $z = -0{,}50$ que, como se viu na Seção 10.2, é 0,3085 ou 3[illegible]35%.

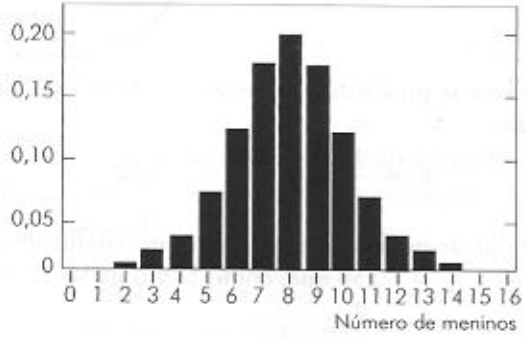
10.4 - Aproximação Normal da Binomial

Considere a variável que representa o número de meninos em 16 nascituros. Essa variável pode assumir qualquer valor inteiro entre zero e 16, inclusive. A distribuição dessa variável está apresentada na Tabela 10.1 e na Figura 10.11.

Tabela 10.1 *Distribuição do número de meninos em 16 nascituros*

Número de meninos	Probabilidade (%)
0	0,00153
1	0,02441
2	0,18311
3	0,85449
4	2,77710
5	6,66504
6	12,21924
7	17,45605
8	19,63806
9	17,45605
10	12,21924
11	6,66504
12	2,77710
13	0,85449
14	0,18311
15	0,02441
16	0,00153

Figura 10.11 *Distribuição do número de meninos em 16 nascituros*



A distribuição binomial apresentada na Figura 10.11 tem configuração semelhante à da distribuição normal. Isto acontece toda vez que $np > 5$ e $nq > 5$. Diz-se então que a distribuição binomial aproxima-se de uma distribuição normal. Essa aproximação é usada para calcular probabilidades associadas aos valores de X.

Por exemplo, suponha que se deseja obter a probabilidade de serem, do sexo masculino, mais de 10 dos 16 recém-nascidos que estão no berçário de determinado hospital. Para fazer os cálculos, pode-se usar a distribuição binomial. Na Tabela 10.1 estão apresentadas todas as probabilidades associadas a uma distribuição binomial com $n = 16$ e $p = \frac{1}{2}$. Fica então fácil obter a probabilidade de serem do sexo masculino mais de 10 dos 16 recém-nascidos. Basta somar as probabilidades de ocorrerem 11, 12, 13, 14, 15 e 16 meninos:

$$6{,}66504 + 2{,}77710 + 0{,}85449 + 0{,}18311 + 0{,}02441 + 0{,}00153 = 10{,}50568\%$$

O cálculo das probabilidades apresentadas na Tabela 10.1 exige, porém, certo trabalho. O problema teria sido mais facilmente resolvido usando a distribuição normal, que pode ser aplicada neste caso porque:

$$np = 16 \cdot \frac{1}{2} = 8 > 5$$

e

$$nq = 16 \cdot \frac{1}{2} = 8 > 5$$

Para resolver o problema usando a distribuição normal é preciso, primeiro, calcular a média e o desvio padrão dessa distribuição. No caso do exemplo tem-se que:

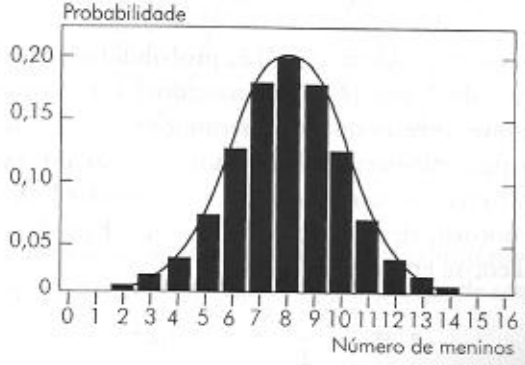
$$\mu = np = 16 \cdot \frac{1}{2} = 8$$

e

$$\sigma = \sqrt{npq} = \sqrt{16 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 2$$

Observe agora a Figura 10.12. A distribuição binomial é discreta e a distribuição normal é contínua. Então, para calcular a probabilidade de serem do sexo masculino mais de 10 recém-nascidos, isto é, 11, 12, 13, 14, 15 ou 16 — toma-se $x = 10{,}5$, em lugar de 10. Esta é a *correção de continuidade*.

Figura 10.12 *Distribuição normal sobre a distribuição binomial*



Agora é preciso calcular:

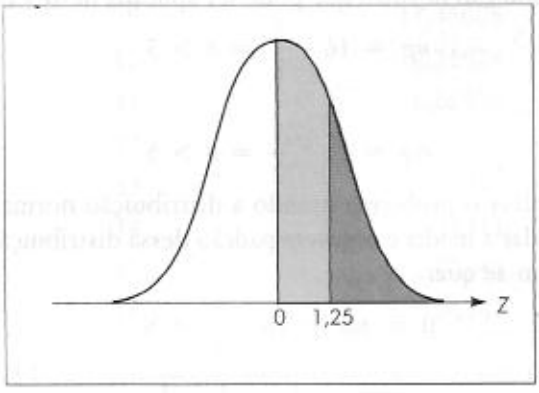
$$z = \frac{10,5 - 8}{2} = 1,25$$

Na tabela de distribuição normal encontra-se, para $z = 1,25$, o valor 0,3944, que corresponde à área pontilhada da Figura 10.13. Então a probabilidade de serem do sexo masculino mais de 10 dos 16 recém-nascidos (área hachurada) é

$$0,5 - 0,3944 = 0,1056,$$

ou seja, 10,56%, valor aproximadamente igual ao achado anteriormente.

Figura 10.13 *Probabilidade de mais de 10 meninos em 16 recém-nascidos*



A correção de continuidade exige um pouco mais de explicação. Na distribuição binomial a variável é discreta e, na normal, a variável é contínua. Então, ao valor $X = k$ da distribuição binomial corresponde o intervalo de $k - 0,5$ a $k + 0,5$ da distribuição normal. Reveja o exemplo. Para obter a probabilidade de X assumir valor maior do que 10 (mais de 10 recém-nascidos), tomou-se $X = 10,5$. Isto porque o primeiro valor inteiro maior do que 10 é 11, que corresponde ao intervalo de 10,5 a 11,5, na distribuição normal.

Veja agora outro problema. Qual é a probabilidade de serem do sexo masculino menos de 4 dos 16 recém-nascidos? Ora, o primeiro valor abaixo de 4 é 3, que corresponde, na distribuição normal, ao intervalo de 2,5 a 3,5. Então, para calcular a probabilidade de serem do sexo masculino menos de 4 dos 16 recém-nascidos, isto é, para calcular $P(X < 4)$ usando a distribuição normal, deve-se tomar $X = 3,5$. Esta é a correção de continuidade. Tem-se então que:

$$z = \frac{3,5 - 8}{2} = -2,25$$

Na tabela de distribuição normal reduzida encontra-se, para $z = 2{,}25$, o valor 0,4878. Esta é a probabilidade de Z assumir valor entre zero e $z = 2{,}25$. A probabilidade de Z assumir valor maior do que $z = 2{,}25$, que é igual à probabilidade de Z assumir valor menor do que $z = -2{,}25$, é:

$$0{,}5000 - 0{,}4878 = 0{,}0122 \text{ ou } 1{,}22\%,$$

bem próximo do valor dado pela binomial que é 1,06%.

10.5 - Exercícios Resolvidos

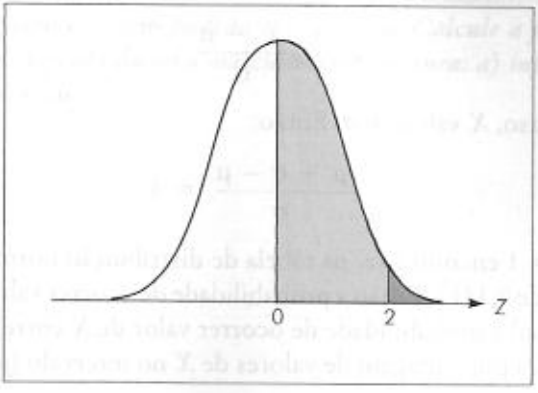
***10.5.1** - Em homens, a quantidade de hemoglobina por 100ml de sangue é uma variável aleatória com distribuição normal de média $\mu = 16g$ e desvio padrão $\sigma = 1g$. Calcule a probabilidade de um homem apresentar de 16 a 18g de hemoglobina por 100ml de sangue.*

Primeiro, é preciso calcular:

$$z = \frac{x - \mu}{\sigma} = \frac{18 - 16}{1} = 2$$

A probabilidade de X assumir valor entre a média 16 e o valor 18 corresponde à probabilidade de Z assumir valor entre a média zero e o valor 2 (área pontilhada na Figura 10.14). Esta probabilidade, dada na tabela, é 0,4772. Então a probabilidade de um homem apresentar de 16 a 18g de hemoglobina por 100ml de sangue é 0,4772 ou 47,72%.

Figura 10.14 *Probabilidade de taxa de hemoglobina entre 16 e 18*



***10.5.2** - No problema anterior, qual é a probabilidade de um homem apresentar mais de 18g de hemoglobina por 100ml de sangue?*

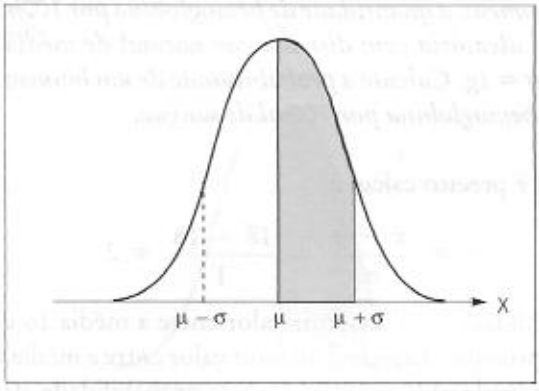
A Figura 10.14 mostra uma área com dupla hachura, que corresponde à probabilidade procurada. Como para $x = 18$ corresponde $z = 2$, e a probabilidade de Z assumir valor entre a média zero e o valor $z = 2$ é 0,4772, segue-se que a probabilidade de Z assumir valor maior do que 2 é:

$$0{,}5 - 0{,}4772 = 0{,}0228 \text{ ou } 2{,}28\%.$$

10.5.3 - *Dada uma variável aleatória X com distribuição normal de média μ e desvio padrão σ, qual é a porcentagem de valores de X compreendidos no intervalo $[\mu - \sigma; \mu + \sigma]$?*

Observe o gráfico apresentado na Figura 10.15.

Figura 10.15 *Gráfico da distribuição normal*



A probabilidade de X assumir valor entre μ e $\mu + \sigma$ corresponde à área pontilhada. Como a distribuição normal é simétrica, essa probabilidade é igual à probabilidade de X assumir valor μ e $\mu - \sigma$. Para utilizar a tabela de distribuição normal reduzida, é necessário obter:

$$Z = \frac{X - \mu}{\sigma}$$

Neste caso, X vale $\mu + \sigma$. Então:

$$z = \frac{\mu + \sigma - \mu}{\sigma} = 1.$$

Para $z = 1$ encontra-se, na tabela de distribuição normal reduzida, a probabilidade 0,3413. Então a probabilidade de ocorrer valor de X entre μ e $\mu + \sigma$ (igual à probabilidade de ocorrer valor de X entre μ e $\mu - \sigma$) é 0,3413. Logo, a porcentagem de valores de X no intervalo $[\mu - \sigma; \mu + \sigma]$ é:

$$34{,}13\% + 34{,}13\% = 68{,}26\%.$$

Este resultado mostra que, se uma variável tem distribuição normal de média μ e desvio padrão σ, cerca de 68% dos valores dessa variável estão compreendidos no intervalo $\mu \pm \sigma$.

10.6 - Exercícios Propostos

10.6.1 - *Suponha que a pressão sangüínea sistólica em indivíduos com idade entre 15 e 25 anos é uma variável aleatória com distribuição aproximadamente normal de média μ = 120mmHg e desvio padrão σ = 8mmHg. Nestas condições, calcule a probabilidade de um indivíduo dessa faixa etária apresentar pressão: a) entre 110 e 130mmHg; b) maior do que 130mmHg.*

10.6.2 - *Suponha que a taxa de glicose no sangue humano é uma variável aleatória com distribuição normal de média μ = 100mg por 100ml de sangue e desvio padrão σ = 6mg por 100ml de sangue. Calcule a probabilidade de um indivíduo apresentar taxa: a) superior a 110mg por 100ml de sangue; b) entre 90 e 100mg por 100ml de sangue.*

10.6.3 - *Suponha que o tempo médio de permanência em um hospital para doenças crônicas sejam 50 dias, com um desvio padrão igual a 10 dias. Se for razoável pressupor que o tempo de permanência tem distribuição aproximadamente normal, qual é a probabilidade de um paciente permanecer no hospital: a) mais de 30 dias? b) menos de 30 dias?*

10.6.4 - *Suponha que a estatura de recém-nascidos do sexo masculino é uma variável aleatória com distribuição aproximadamente normal de média μ = 50cm e desvio padrão σ = 2,50cm. Calcule a probabilidade de um recém-nascido do sexo masculino ter estatura: a) inferior a 48cm; b) superior a 52cm.*

CAPÍTULO
11

Teste de χ^2

Muitas vezes o pesquisador toma decisão para o todo (população), tendo examinado apenas parte (amostra). Esse processo chama-se *inferência*. Na pesquisa científica a inferência é feita com a ajuda de testes estatísticos. Mas o que são testes estatísticos? Para responder a esta pergunta é preciso, primeiro, firmar alguns conceitos. Isto será feito através de um exemplo.

11.1 - Conceitos Básicos

A proporção de recém-nascidos com defeito ou doença séria é 3%. Imagine que um médico suspeita que esta proporção aumentou. Para estabelecer se a suspeita é procedente, é preciso fazer um teste estatístico.

São duas as *hipóteses* em teste: a primeira é a de que a proporção de recém-nascidos com defeito ou doença séria continua igual a 3%; a segunda é a de que a proporção de recém-nascidos com defeito ou doença séria é maior do que 3%.

O estatístico chama a primeira hipótese de *hipótese da nulidade* e a indica por H_0 (lê-se agá-zero). Chama a segunda hipótese de *hipótese alternativa* e a indica por H_1 (lê-se agá-um). Escreve:

$$H_0: p = 0{,}03$$
$$H_1: p > 0{,}03$$

Estabelecidas as hipóteses, é preciso obter, num grande número de recém-nascidos, a proporção de portadores de defeito ou doença séria. Por exemplo, é razoável examinar 1 000 recém-nascidos para determinar o número de portadores de defeito ou doença séria.

Se, dos 1 000 recém-nascidos, 30 ou menos apresentarem defeito ou doença séria, é razoável concluir que a suspeita do médico é infundada. Mas se ocorrerem 31? Ou 32? Ou mesmo 33 ou 34? Então é preciso estabelecer um número k, a partir do qual se passa a admitir que o médico tem razão quando suspeita que a proporção de recém-nascidos com defeito ou doença séria aumentou.

Seja $k = 40$, isto é, a hipótese de que a proporção de recém-nascidos com defeito ou doença séria é 3% deve ser rejeitada se, na amostra de 1 000 recém-nascidos, ocorrerem 40 ou mais portadores de defeito ou doença séria. Então foi estabelecida uma "regra de decisão". Será que esta regra de decisão é correta?

Toda inferência (decisão para o todo com base no exame de parte) está sujeita a erro. Decidiu-se rejeitar H_0 se, na amostra de 1 000 recém-nascidos, 40 ou mais apresentarem defeito ou doença séria. Mas é possível — mesmo que a proporção de portadores de defeito ou doença séria seja 3% — que numa amostra de 1 000 recém-nascidos ocorram 40 ou mais portadores de defeito ou doença séria. Qual é a probabilidade de isto acontecer?

O número de portadores de defeito ou doença séria em 1 000 recém-nascidos é uma variável aleatória com distribuição binomial. Se H_0 for verdadeira, essa distribuição tem média

$$\mu = np = 1\,000 \cdot 0{,}03 = 30$$

e desvio padrão

$$\sigma = \sqrt{npq} = \sqrt{1000 - 0{,}03 \cdot 0{,}97} = 5{,}39.$$

Toda vez que $np > 5$ e $nq > 5$, a distribuição binomial aproxima-se de uma distribuição normal. No exemplo, $np = 1\,000 \cdot 0{,}03 = 30$ e $nq = 1\,000 \cdot 0{,}97 = 970$. Então o número de portadores de defeito ou doença séria em 1 000 recém-nascidos é uma variável aleatória com distribuição aproximadamente normal. Nestas condições, qual é a probabilidade de ocorrerem 40 ou mais portadores de defeito ou doença séria, numa amostra de 1 000 recém-nascidos?

Lembre que é preciso calcular:

$$z = \frac{(x - 0{,}5) - \mu}{\sigma}$$

$$= \frac{(40 - 0{,}5) - 30}{5{,}39}$$

$$= 1{,}76$$

A tabela de distribuição normal reduzida dá, para $z = 1{,}76$, o valor 0,4608. Então a probabilidade de ocorrerem 40 ou mais recém-nascidos com defeito ou doença séria, em 1 000 recém-nascidos, é:

$$0{,}5000 - 0{,}4608 = 0{,}0392 \text{ ou } 3{,}92\%.$$

Em termos da regra de decisão, o que isto significa? Ora, decidiu-se rejeitar a hipótese de que a proporção de recém-nascidos com defeito ou doença séria é 3% se, em 1 000 recém-nascidos, ocorrerem 40 ou mais portadores de defeito ou doença séria. Mas — mesmo que H_0 seja verdadeira — podem ocorrer 40 ou mais portadores de defeito ou doença séria em 1 000 recém-nascidos. A probabilidade é 3,92%. Então existe uma probabilidade de 3,92% de rejeitar H_0, quando H_0 é verdadeira.

Essa probabilidade — de rejeitar H_0, quando H_0 é verdadeira — é o *nível de significância* do teste. O estatístico não sabe, quando rejeita H_0, se está ou não cometendo erro, mas sabe a probabilidade de cometer esse tipo de erro. Se essa probabilidade for suficientemente pequena, como no caso do exemplo (3,92%), a decisão de rejeitar H_0 está bem fundamentada. É usual representar o nível de significância de um teste pela letra grega α (lê-se alfa).

11.2 - Procedimentos Usuais

O pesquisador levanta dados para responder a uma pergunta. O estatístico transforma a pergunta do pesquisador em hipóteses. Por exemplo, se o pesquisador pergunta: "Será que a droga A cura tanto quanto a droga B?", o estatístico vê a pergunta do pesquisador como duas hipóteses:

H_0: a proporção de pacientes curados com a droga A é igual à proporção de pacientes curados com a droga B.

H_1: a proporção de pacientes curados com a droga A é diferente da proporção de pacientes curados com a droga B.

Feitas as hipóteses, o estatístico estabelece o *nível de significância* do teste. No caso deste exemplo, o nível de significância seria a probabilidade de afirmar que uma das drogas determina maior proporção de curas quando, na verdade, a proporção de pacientes curados é a mesma, quer se use A ou B. É usual manter o nível de significância em $\alpha = 1\%$, $\alpha = 5\%$ ou $\alpha = 10\%$.

Estabelecido o nível de significância, o estatístico escolhe o teste apropriado. Existe hoje grande variedade de testes à disposição dos interessados. Todos têm indicação precisa e todos têm vantagens e desvantagens. Então a escolha do teste exige conhecimento de estatística.

Feito o teste, chega-se a um valor numérico e, com base nesse valor, decide-se se a hipótese da nulidade deve ser rejeitada ao nível de significância estabelecido. O pesquisador deve, então, discutir esta informação.

11.3 - Teste de χ^2 para Aderência

Um pesquisador pode ter interesse em verificar se a distribuição dos elementos, numa população, está de acordo com uma dada teoria. O exemplo que será usado aqui pertence à história da ciência e constitui a base da Genética.

Em 1866, o monge austríaco Gregor Mendel relatou os resultados de seus trabalhos na hibridação de ervilhas. Em um célebre experimento Mendel polinizou 15 plantas de sementes lisas e albume amarelo com plantas de sementes rugosas e albume verde. As plantas resultantes desse cruzamento tinham sementes lisas e albume amarelo (amarelo-lisas). Cruzando essas plantas entre si, Mendel obteve 556 sementes, distribuídas conforme mostra a Tabela 11.1.

Tabela 11.1

Distribuição das ervilhas em um dos experimentos de Mendel

Sementes	Freqüência
Amarelo-lisas	315
Amarelo-rugosas	101
Verde-lisas	108
Verde-rugosas	32
Total	556

Fonte: BISHOP, FIENBERG e HOLLAND (1975)

A teoria postulada por Mendel estabelece que a segregação, neste caso, deve ocorrer na seguinte proporção:

$$\frac{9}{16}:\frac{3}{16}:\frac{3}{16}:\frac{1}{16}$$

Será que os resultados obtidos experimentalmente por Mendel estão de acordo com a teoria que ele postulava? Ora, foram obtidas 556 ervilhas. Então a *freqüência esperada* de amarelo-lisas é:

$$\frac{9}{16} \cdot 556 = 312{,}75,$$

a *freqüência esperada* de amarelo-rugosas é:

$$\frac{3}{16} \cdot 556 = 104{,}25,$$

a *freqüência esperada* de verde-lisas é:

$$\frac{3}{16} \cdot 556 = 104{,}25$$

e a *freqüência esperada* de verde-rugosas é:

$$\frac{1}{16} \cdot 556 = 34{,}75.$$

Todos estes valores estão apresentados na Tabela 11.2.

Tabela 11.2

Distribuição dos valores esperados no experimento de Mendel

Sementes	Freqüência
Amarelo-lisas	312,75
Amarelo-rugosas	104,25
Verde-lisas	104,25
Verde-rugosas	34,75
Total	556,00

Compare a Tabela 11.1 com a Tabela 11.2. As diferenças entre as freqüências observadas e esperadas são, respectivamente:

$$\begin{aligned} 315 - 312{,}75 &= 2{,}25 \\ 101 - 104{,}25 &= -3{,}25 \\ 108 - 104{,}25 &= 3{,}75 \\ 32 - 34{,}75 &= -2{,}75 \end{aligned}$$

Para verificar se a distribuição de freqüências observadas está de acordo com a teoria, aplica-se um *teste de aderência*. O mais conhecido desses testes é o teste de χ^2 (lê-se qui-quadrado), apresentado aqui. Para aplicar este teste é preciso:

a) estabelecer o nível de significância
b) calcular o valor de qui-quadrado, dado pela fórmula:

$$\chi^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i}$$

onde O_i $(i = 1.....r)$ representa as freqüências observadas e E_i representa as freqüências esperadas;

c) comparar o valor calculado de χ^2 com o valor da tabela, ao nível de significância estabelecido e com $r - 1$ graus de liberdade. Toda vez que o valor calculado de χ^2 for igual ou maior do que o valor da tabela, rejeita-se a hipótese de que a distribuição das freqüências observadas está de acordo com a teoria, ao nível de significância estabelecido.

O teste de χ^2 pode ser aplicado aos dados obtidos experimentalmente por Mendel. Para isso, é preciso:

a) estabelecer o nível de significância.

Seja $\alpha = 5\%$.

b) calcular o valor de qui-quadrado:

$$\chi^2 = \frac{2{,}25^2}{312{,}75} + \frac{(-3{,}25)^2}{104{,}25} + \frac{3{,}75^2}{104{,}25} + \frac{(-2{,}75)^2}{34{,}75} = 0{,}47$$

c) comparar o valor calculado de χ^2 com o valor da tabela de χ^2, ao nível de significância de 5% e com $4 - 1 = 3$ graus de liberdade.

Para entender como se usa a tabela de χ^2, observe a Figura 11.1, que reproduz parte da Tabela A.2, apresentada neste livro, em Apêndice. Foi sombreado o valor de χ^2 com 3 graus de liberdade, ao nível de significância de 5%.

Figura 11.1 *Valor de χ^2 para $\alpha = 5\%$ e com três graus de liberdade*

Valores de χ^2, segundo os graus de liberdade e o valor de α

Graus de liberdade	α 10%	5%	1%
1	2,71	3,84	6,64
2	4,60	5,99	9,21
3	6,25	7,82	11,34
4	7,78	9,49	13,28
5	9,24	11,07	15,09

O valor de χ^2 calculado com base no célebre experimento de Mendel ($\chi^2 = 0{,}47$) é menor do que o valor dado em tabela ($\chi^2 = 7{,}82$). Então não se rejeita, ao nível de significância de 5%, a hipótese de que a segregação ocorreu de acordo com a teoria.

11.4 - Teste de χ^2 para Independência

Um pesquisador pode ter interesse em verificar se duas populações têm a mesma proporção de indivíduos com determinada característica. Por exemplo, será que a proporção de natimortos é a mesma nos dois sexos?

A Tabela 11.3 apresenta a proporção de natimortos, segundo o sexo. Será que a diferença das proporções é suficientemente grande para permitir rejeitar a hipótese de que a proporção de natimortos é a mesma nos dois sexos?

Tabela 11.3

Recém-nascidos segundo o sexo e a condição de vivo ou natimorto

Sexo	Condição		Proporção de natimortos
	Vivo	Natimorto	
Masculino	1 513	37	2,39
Feminino	1 451	27	1,83

Fonte: ARENA (1977)

Para testar a hipótese de nulidade, isto é, a hipótese de que a proporção de natimortos é a mesma nos dois sexos, aplica-se o teste de χ^2. Mas é preciso, primeiro, estabelecer o nível de significância. Seja $\alpha = 5\%$. Depois, é preciso comparar as freqüências observadas com as freqüências esperadas sob a hipótese de nulidade. Mas como se calculam essas freqüências?

Se a hipótese da nulidade fosse verdadeira, isto é, se a proporção de natimortos fosse a mesma nos dois eixos, quantos natimortos seriam do sexo masculino e quantos seriam do sexo feminino? Para responder a esta pergunta é preciso, primeiro, calcular os totais apresentados na Tabela 11.4.

Tabela 11.4

Recém-nascidos segundo o sexo e a condição de vivo ou natimorto

Sexo	Condição		Proporção de natimortos
	Vivo	Natimorto	
Masculino	1 513	37	1 550
Feminino	1 451	27	1 478
Total	2 964	64	3 028

Fonte: ARENA (1977)

Do total de 3 028 recém-nascidos, 2 964 nasceram vivos. Então a freqüência esperada de meninos nascidos vivos entre os 1 550 meninos é E_{11}, tal que:

$$3028 \rightarrow 2964$$
$$1550 \rightarrow E_{11}$$

Portanto;

$$E_{11} = \frac{1550 \cdot 2964}{3028} = 1517{,}24$$

Dos 3 028 recém-nascidos, 64 nasceram mortos. A *freqüência esperada* de natimortos do sexo masculino entre os 1 550 recém-nascidos de sexo masculino é E_{12}, tal que:

$$3028 \rightarrow 64$$
$$1550 \rightarrow E_{12}$$

Portanto

$$E_{12} = \frac{1550 \cdot 64}{3028} = 32{,}76$$

Dos 3 028 recém-nascidos, 2 964 nasceram vivos. Como 1 478 são meninas, a *freqüência esperada* de meninas nascidas vivas é E_{21}, tal que:

$$3028 \rightarrow 2964$$
$$1478 \rightarrow E_{21}$$

Logo

$$E_{21} = \frac{1478 \cdot 2964}{3028} = 1446{,}76$$

Dos 3 028 recém-nascidos, 64 nasceram mortos. Como 1 478 são meninas, a *freqüência esperada* de natimortos do sexo feminino é E_{22}, tal que:

$$3028 \rightarrow 64$$
$$1478 \rightarrow E_{22}$$

Logo

$$E_{22} = \frac{1478 \cdot 64}{3028} = 31{,}24$$

As freqüências esperadas estão apresentadas na Tabela 11.5. Note que, calculado o valor de E_{11}, as demais freqüências esperadas podem ser obtidas por diferença.

Tabela 11.5

Valores esperados de recém-nascidos vivos e natimortos segundo o sexo

Sexo	Condição		Total
	Vivo	Natimorto	
Masculino	1 517,24	32,76	1550
Feminino	1 446,76	31,24	1478
Total	2 964,00	64,00	3028

Não se rejeita a hipótese da nulidade quando as freqüências esperadas são iguais às freqüências observadas. Mas compare as Tabelas 11.4 e 11.5. As diferenças entre as freqüências observadas e as freqüências esperadas são as seguintes:

$$\begin{aligned} 1513 - 1517{,}24 &= -4{,}24 \\ 37 - 32{,}76 &= 4{,}24 \\ 1451 - 1446{,}76 &= 4{,}24 \\ 27 - 31{,}24 &= -4{,}24 \end{aligned}$$

Será que essas diferenças são suficientemente grandes para que se possa rejeitar a hipótese de nulidade? Para responder a esta questão é preciso calcular o valor de χ^2, dado pela fórmula:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

As tabelas de contingência r x s, isto é, com r linhas e s colunas, estão associadas a $(r-1)\,(s-1)$ graus de liberdade. A tabela de contingência do exemplo é 2 x 2 (dois sexos, duas condições). Então essa tabela está associada a $(2-1)\,(2-1) = 1$ grau de liberdade.

O valor de χ^2 é:

$$\chi^2 = \frac{(-4{,}24)^2}{1517{,}24} + \frac{4{,}24^2}{32{,}76} + \frac{4{,}24^2}{1446{,}76} + \frac{(-4{,}24)^2}{31{,}24} = 1{,}15$$

Toda vez que o valor calculado de χ^2 é igual ou maior do que o valor dado na tabela, rejeita-se H_0 ao nível de significância estabelecido. Na Tabela A.2 do Apêndice, de χ^2, para o nível de significância de 5% e com 1 grau de liberdade, encontra-se o valor 3,84. Como o valor calculado de χ^2 (1,15) é menor do que 3,84, não se rejeita a hipótese de que a proporção de natimortos é a mesma nos dois sexos.

11.5 - Restrições ao Uso do Teste de χ^2

Por razões teóricas, o teste de χ^2 tem as seguintes restrições:

a) só deve ser aplicado quando a amostra tem mais de 20 elementos;
b) quando a amostra tem mais de 20, mas menos de 40 elementos, só deve ser aplicado se todas as freqüências esperadas forem maiores do que 1.

Finalmente, o teste de χ^2 é aproximado. A aproximação melhora bastante quando se faz a correção de continuidade. No caso das tabelas 2 x 2, esta correção, conhecida como correção de Yates, consiste em calcular o valor de χ^2 através da fórmula:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(|O_{ij} - E_{ij}| - 0{,}5)^2}{E_{ij}}$$

11.6 - Risco Relativo

A Seção 8.2 do Capítulo 8 apresenta a definição de probabilidade condicional. Mas probabilidade condicional também pode ser estimada com base em amostras. Veja a Tabela 11.6. Os dados apresentados permitem estimar a probabilidade condicional de um nascituro ter aberração cromossômica, dado que a gestante está na faixa etária de 35 até 40 anos, e a probabilidade (condicional) de um nascituro ter aberração cromossômica, dado que a gestante tem 40 anos ou mais.

Tabela 11.6

Resultados de casos de diagnóstico pré-natal segundo a idade da gestante e a presença ou ausência de aberração cromossômica.

Idade da gestante	Aberração cromossômica		Total
	Presente	Ausente	
de 35 até 40	10	447	457
40 ou mais	18	510	528

Fonte: MILUNSKY e ATKINS (1977)

Na área de saúde, é comum usar a palavra *risco* para identificar a probabilidade de um evento indesejável. Então, com base nos dados da Tabela 11.6, é possível obter:

a) o *risco* de um nascituro ter aberração cromossômica, *dado* que a gestante está na faixa etária de 35 até 40 anos:

$$\frac{10}{457} = 0{,}0219 \text{ ou } 2{,}19\%$$

b) o risco de um nascituro ter aberração cromossômica, *dado* que a gestante tem 40 anos ou mais:

$$\frac{18}{528} = 0{,}0341 \text{ ou } 3{,}41\%$$

Risco relativo é a razão entre duas probabilidades condicionais (ou dois riscos condicionais). Com base na Tabela 11.6, tem-se o risco relativo

$$\frac{3{,}41}{2{,}19} = 1{,}56$$

Este resultado mostra que o risco de um nascituro apresentar aberração cromossômica é 1,56 maior se a gestante tiver 40 anos ou mais do que se a gestante estiver na faixa etária de 35 até 40 anos.

Veja outro exemplo. Com base nos dados apresentados na Tabela 11.7, é possível estimar o risco de um recém-nascido ser defeituoso, dada a época do ataque de rubéola na gestante.

Tabela 11.7

Recém-nascidos segundo a época do ataque de rubéola na gestante e a condição

Época do Ataque	Condição		Total
	Normal	Defeituoso	
Até o 3º mês	36	14	50
Depois do 3º mês	51	3	54
Total	87	17	104

Fonte: HILL et alii (1958)

O risco de um recém-nascido ser defeituoso, dado que o ataque de rubéola na gestante ocorreu no primeiro trimestre de gestação, é

$$\frac{14}{50} = 0{,}28 \text{ ou } 28\%$$

e dado que o ataque ocorreu depois do 3º mês de gestação, é

$$\frac{3}{54} = 0{,}0556 \text{ ou } 5{,}56\%$$

O risco relativo é

$$r = \frac{0{,}28}{0{,}0556} = 5{,}04,$$

ou seja, é aproximadamente 5 vezes mais provável que um recém-nascido tenha defeito se o ataque de rubéola na gestante ocorreu no primeiro trimestre de gestação. Consagrou-se a expressão *grupo de risco* para identificar o grupo de maior risco. No caso deste exemplo, nascituros de mães que tiveram rubéola até o 3º mês de gestação estão no grupo de risco de ter defeito.

Para testar a hipótese de que o risco relativo é significante, isto é, para testar a hipótese de que o risco relativo é 1, contra a hipótese de que é maior do que 1, aplica-se o teste de χ^2, na forma apresentada neste capítulo.

11.7 - Medidas de Associação

Muitas vezes existe interesse em medir o *grau de associação* de duas variáveis qualitativas. Por exemplo, pode haver interesse em verificar se a incidência de determinada doença está associada a sexo. Para medir o grau de associação de duas variáveis qualitativas, usam-se os *coeficientes de associação*. Nesta seção será explicado o coeficiente de Yule, que só se

aplica às tabelas 2 x 2. Mas existem coeficientes de associação para tabelas $r \times s$, onde $r > 2$ e $s > 2$, como é o caso do coeficiente de Tschuprow.

Para entender o que é uma associação entre variáveis, veja a Tabela 11.8.

A Tabela 11.8 mostra que 12 cobaias receberam xilocaína (anestésico local) com pH 4,7, e 12 cobaias receberam xilocaína com pH 7,4. A proporção de cobaias anestesiadas com pH 4,7 foi 8/12 e de cobaias anestesiadas com pH 7,4 foi 3/12. Então a eficiência do anestésico aumenta quando diminui o pH do anestésico. Isto significa que as variáveis estão associadas e a associação entre elas é negativa (porque à medida que uma aumenta, a outra diminui).

Tabela 11.8

Cobaias segundo o pH da solução de xilocaína e a sensibilidade

pH	Sensibilidade		Total
	Sim	Não	
4,7	4	8	12
7,4	9	3	12

Fonte: GAMA (1976)

Para entender como se calcula o coeficiente de Yule, veja a Tabela 11.9, que apresenta os termos genéricos de uma Tabela 2 x 2.

Tabela 11.9

Elementos classificados segundo a variável A e a variável B

Variável A	Variável B	
	B_1	B_2
A_1	n_{11}	n_{12}
A_2	n_{21}	n_{22}

O coeficiente de Yule é definido pela fórmula:

$$Y = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}$$

O coeficiente de Yule varia entre –1 e +1. Se a associação entre as variáveis é negativa, o coeficiente de Yule é negativo. Se a associação entre as variáveis é positiva, o coeficiente de Yule é positivo.

Para os dados da Tabela 11.8, o coeficiente de Yule é:

$$Y = \frac{4.3 - 9.8}{4.3 + 9.8} = -\frac{60}{84} = -0{,}71$$

o que significa que a proporção de cobaias anestesiadas aumenta quando se diminui o pH do anestésico. O teste de χ^2, apresentado neste capítulo, mostra se a associação entre as variáveis é significante.

11.8 - Exercícios Resolvidos

11.8.1 - ***Com base nos dados apresentados na Tabela 11.7, teste a hipótese de que a proporção de recém-nascidos defeituosos é a mesma, qualquer que tenha sido a época em que a gestante foi atacada de rubéola.***

Seja $\alpha = 1\%$. Para calcular o valor de χ^2 é preciso, primeiro, obter as freqüências esperadas sob a hipótese de nulidade, isto é, sob a hipótese de que a proporção de recém-nascidos defeituosos é a mesma, qualquer que tenha sido a época em que a gestante foi atacada de rubéola.

Então, se dos 104 recém-nascidos 87 eram normais, dos 50 filhos de mães que tiveram rubéola até o 3º mês de gestação, esperam-se E_{11} normais, tal que:

$$104 \rightarrow 87$$
$$50 \rightarrow E_{11}$$

Portanto,

$$E_{11} = \frac{50 \cdot 87}{104} = 41{,}83$$

Da mesma forma, se dos 104 recém-nascidos 17 eram defeituosos, dos 50 filhos de mães que tiveram rubéola até o 3º mês de gestação esperam-se E_{12} defeituosos, tal que:

$$104 \rightarrow 17$$
$$50 \rightarrow E_{12}$$

Portanto,

$$E_{12} = \frac{50 \cdot 17}{104} = 8{,}17$$

Raciocinando da mesma forma, obtém-se a freqüência esperada de recém-nascidos normais filhos de mães que tiveram rubéola depois do 3º mês de gestação.

$$104 \rightarrow 87$$
$$54 \rightarrow E_{21}$$

Então,

$$E_{21} = \frac{54 \cdot 87}{104} = 45{,}17$$

e de recém-nascidos defeituosos filhos de mães que tiveram rubéola depois do 3º mês de gestação:

$$104 \rightarrow 17$$
$$54 \rightarrow E_{22}$$

Então

$$E_{22} = \frac{54 \cdot 17}{104} = 8{,}83$$

De posse das freqüências esperadas, calcula-se:

$$\chi^2 = \frac{(36-41{,}83)^2}{41{,}83} + \frac{(14-8{,}17)^2}{8{,}17} + \frac{(51-45{,}17)^2}{45{,}17} + \frac{(3-8{,}83)^2}{8{,}83} =$$

$$= 0{,}81 + 4{,}16 + 0{,}75 + 3{,}85 = 9{,}57$$

que está associado a 1 grau de liberdade. Na Tabela A.2 do Apêndice, de χ^2, para $\alpha = 1\%$ e com 1 grau de liberdade, tem-se o valor 6,64. Como o valor calculado é maior do que 6,64, conclui-se que a proporção de recém-nascidos defeituosos é maior quando o ataque de rubéola na gestante ocorre nos três primeiros meses de gestação.

11.8.2 - *Com base nos dados apresentados na Tabela 11.10, proceda ao teste de χ^2, ao nível de significância de 1%, para testar a hipótese de que o tipo sangüíneo independe da origem do indivíduo.*

Tabela 11.10

Indivíduos segundo a origem e o tipo sangüíneo

Origem	Tipo sangüíneo				Total
	O	A	B	AB	
Árabe	130	149	29	8	316
Não-árabe	417	292	94	17	820
Total	547	441	123	25	1136

Fonte: GARCIA (1977)

Para proceder ao teste de χ^2 é preciso, primeiro, obter as freqüências esperadas sob a hipótese da nulidade, isto é, sob a hipótese de que a proporção de indivíduos com cada tipo sangüíneo não depende da origem. Então é preciso calcular:

$$E_{11} = \frac{547 \cdot 316}{1136} = 152{,}16$$

$$E_{12} = \frac{441 \cdot 316}{1136} = 122{,}67$$

$$E_{13} = \frac{123 \cdot 316}{1136} = 34{,}21$$

$$E_{14} = \frac{25 \cdot 316}{1136} = 6{,}95$$

$$E_{21} = \frac{547 \cdot 820}{1136} = 394{,}84$$

$$E_{22} = \frac{441 \cdot 820}{1136} = 318{,}33$$

$$E_{23} = \frac{123 \cdot 820}{1136} = 88{,}79$$

$$E_{24} = \frac{25 \cdot 820}{1136} = 18{,}05$$

Agora é preciso calcular:

$$\chi^2 = \frac{(130 - 152{,}16)^2}{152{,}16} + \frac{(149 - 122{,}67)^2}{122{,}67} + \frac{(29 - 34{,}21)^2}{34{,}21} +$$

$$+ \frac{(8 - 6{,}95)^2}{6{,}95} + \frac{(417 - 394{,}84)^2}{394{,}84} + \frac{(292 - 318{,}33)^2}{318{,}33} +$$

$$+ \frac{(94 - 88{,}79)^2}{88{,}79} + \frac{(17 - 18{,}05)^2}{18{,}05}$$

$$= 3{,}23 + 5{,}65 + 0{,}79 + 0{,}16 + 1{,}24 + 2{,}18 + 0{,}31 + 0{,}06 = 13{,}62$$

Este valor está associado a $(2-1)(4-1) = 3$ graus de liberdade. A Tabela A.2 do Apêndice dá, para $\alpha = 1\%$ e com 3 graus de liberdade, o valor 11,34. Como o valor calculado é maior do que o valor da tabela, conclui-se que o tipo sangüíneo depende da origem do indivíduo.

11.9 - Exercícios Propostos

11.9.1 - Com base nos dados apresentados na Tabela 11.11 teste, ao nível de significância de 5%, a hipótese de que a proporção de recém-nascidos vivos portadores de anomalia é a mesma nos dois sexos.

Tabela 11.11

Recém-nascidos vivos segundo o sexo e a presença ou ausência de anomalia

Sexo	Anomalia		Total
	Presente	Ausente	
Masculino	28	1 485	1 513
Feminino	45	1 406	1 451
Total	73	2 891	2 964

Fonte: ARENA (1977)

11.9.2 - *Com base nos dados apresentados na Tabela 11.12 teste, ao nível de significância de 5%, a hipótese de que a proporção de pessoas com Rh^- não depende da origem.*

Tabela 11.12

Indivíduos segundo a origem e o fator Rh

Origem	Fator	
	Rh^+	Rh^-
Árabe	289	27
Não-árabe	737	83

Fonte: GARCIA (1977)

11.9.3 - *Com base nos dados apresentados na Tabela 11.13 teste, ao nível de significância de 1%, a hipótese de que a ausência congênita de dentes independe do sexo.*

Tabela 11.13

Escolares segundo o sexo e a ausência congênita de dentes

Sexo	Ausência congênita de dentes	
	Portador	Não-portador
Masculino	23	1 078
Feminino	40	859

Fonte: VEDOVELO FILHO (1972)

11.9.4 - *Com base nos dados apresentados na Tabela 11.5 calcule o risco relativo.*

CAPÍTULO
12

Teste *t*

Às vezes, é preciso comparar duas populações. Por exemplo, imagine que um pesquisador obteve, para um grande número de crianças, a idade em que cada uma delas começou a falar. Para verificar se meninos e meninas aprendem a falar na mesma idade, o pesquisador terá que comparar os dados dos dois sexos.

Outras vezes, é preciso comparar condições experimentais. Por exemplo, para saber se um tratamento tem efeito, organizam-se dois grupos de unidades: um grupo recebe o tratamento em teste (é o *grupo tratado*), enquanto o outro não recebe o tratamento (é o *grupo controle*). O efeito do tratamento é dado pela comparação dos dois grupos.

12.1 - Teste *t* para Observações Independentes

Se a variável em análise tem distribuição normal ou aproximadamente normal, aplica-se o teste *t* para comparar duas médias. Mas primeiro é preciso estabelecer o nível de significância, que se indica pela letra grega α. Depois, dados os dois grupos, 1 e 2, calculam-se:

a) a média de cada grupo; indica-se:

$\bar{x}_1$: média do grupo 1
$\bar{x}_2$: média do grupo 2

b) a variância de cada grupo; indica-se:

s_1^2: variância do grupo 1
s_2^2: variância do grupo 2

c) a variância ponderada, dada pela fórmula:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

onde n_1 é número de elementos do grupo 1 e n_2 é número de elementos do grupo 2.

d) o valor de t, definido por

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Feitos os cálculos, é preciso comparar o valor calculado de t com o valor de uma tabela de t, ao nível de significância estabelecido e com $(n_1 + n_2 - 2)$ graus de liberdade. Para entender como se acha esse valor, observe a Figura 12.1, que apresenta parte da tabela de t dada neste livro, em Apêndice. O valor de t, para o nível de significância de 1% e com 5 graus de liberdade, foi sombreado. Toda vez que o valor calculado de t, em valor absoluto, for igual ou maior do que o valor da tabela, conclui-se que as médias não são iguais, ao nível de significância estabelecido.

Figura 12.1 *Valor de t para α = 1% e 5 graus de liberdade*

Valores de t, segundo os graus de liberdade e o valor de α			
Graus de Liberdade	α 10%	5%	1%
1	6,31	12,71	63,66
2	2,92	4,30	9,92
3	2,35	3,18	5,84
4	2,13	2,78	4,60
5	2,02	2,57	4,03

12.2 - Exemplo de Aplicação

Para verificar se duas dietas para emagrecer são igualmente eficientes, um médico separou, ao acaso, um conjunto de pacientes em dois

grupos. Cada paciente seguiu a dieta designada para seu grupo. Decorrido certo tempo, o médico obteve a perda de peso, em quilogramas, de cada paciente de cada grupo. Os dados estão apresentados na Tabela 12.1.

Tabela 12.1

Perdas de peso, em quilogramas, segundo a dieta

Dieta	
1	2
12	15
8	19
15	15
13	12
10	13
12	16
14	15
11	
12	
13	

Para proceder ao teste t é preciso, primeiro, estabelecer o nível de significância. Seja $\alpha = 5\%$. Depois, é preciso calcular:

a) a média de cada grupo

$$\bar{x}_1 = \frac{12+8+...+13}{10} = \frac{120}{10} = 12$$

$$\bar{x}_2 = \frac{15+19+...+15}{7} = \frac{105}{7} = 15$$

b) a variância de cada grupo

$$s_1^2 = \frac{1476 - \frac{120^2}{10}}{9} = \frac{36}{9} = 4$$

$$s_2^2 = \frac{1605 - \frac{105^2}{7}}{6} = \frac{30}{6} = 5$$

c) a variância ponderada

$$s^2 = \frac{9 \cdot 4 + 6 \cdot 5}{9+6} = 4{,}4$$

d) o valor de t

$$t = \frac{15 - 12}{\sqrt{4{,}4\left(\frac{1}{10} + \frac{1}{7}\right)}} = 2{,}902$$

que está associado a $n_1 + n_2 - 2 = 10 + 7 - 2 = 15$ graus de liberdade. Na Tabela A.6 do Apêndice, ao nível de significância de 5% e com 15 graus de liberdade, o valor de t é 2,13. Como o valor calculado (2,902) é maior do que o valor da tabela (2,13), conclui-se que, em média, as perdas de peso de pacientes submetidos aos dois tipos de dieta são diferentes. Em termos práticos, a perda de peso é maior quando os pacientes são submetidos à dieta 2.

12.3 - Teste t para Observações Pareadas

Para estudar o efeito de um tratamento, muitas vezes comparam-se pares de indivíduos. Por exemplo, em alguns estudos de psicologia comparam-se pares de gêmeos: um dos gêmeos recebe o tratamento, enquanto o outro permanece sem o tratamento (controle).

Outras vezes, comparam-se os dois lados dos mesmos indivíduos. Por exemplo, para estudar o efeito de um tratamento para prevenção de cáries, o dentista pode aplicar o tratamento em um lado da arcada dentária de cada paciente, e deixar o outro lado sem tratamento (controle).

Também são feitos experimentos em que se observam os mesmos indivíduos duas vezes, isto é, uma vez antes, outra vez depois de administrar o tratamento. Por exemplo, para verificar o efeito de um tratamento sobre a pressão arterial, o médico pode obter a pressão arterial de seus pacientes, antes e depois de administrar o tratamento.

Todos esses exemplos são de *observações pareadas* (pares de gêmeos, dois lados de um indivíduo, duas observações no mesmo indivíduo). Para testar o efeito de um tratamento, quando as observações são pareadas, aplica-se o teste t. Mas é preciso, primeiro, estabelecer o nível de significância do teste. Depois, é preciso calcular:

a) a diferença entre as unidades de cada um dos n pares

$$d = x_2 - x_1$$

b) a média das diferenças

$$\bar{d} = \frac{\sum d}{n}$$

c) a variância das diferenças

$$s^2 = \frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}$$

d) o valor de t

$$t = \frac{\bar{d}}{\sqrt{\frac{s^2}{n}}}$$

que está associado a $n - 1$ graus de liberdade.

Feitos os cálculos, é preciso procurar o valor de t na tabela, ao nível de significância estabelecido e com $n - 1$ graus de liberdade. Toda vez que o valor absoluto de t calculado for igual ou maior do que o valor da tabela, conclui-se que o tratamento tem efeito ao nível de significância estabelecido.

12.4 - Exemplo de Aplicação

Na Tabela 12.2 são dados os pesos de 9 pessoas, antes e depois da dieta para emagrecimento.

Tabela 12.2

Pesos em quilogramas de 9 pessoas antes e depois da dieta para emagrecimento.

Dieta	
Antes	Depois
77	80
62	58
61	61
80	76
90	79
72	69
86	90
59	51
88	81

Para fazer o teste, é preciso primeiro estabelecer o nível de significância. Seja $\alpha = 1\%$. Depois, é preciso calcular:

a) as diferenças entre os valores observados antes e depois da dieta

$$
\begin{aligned}
80 - 77 &= 3 \\
58 - 62 &= -4 \\
61 - 61 &= 0 \\
76 - 80 &= -4 \\
79 - 90 &= -11 \\
69 - 72 &= -3 \\
90 - 86 &= 4 \\
51 - 59 &= -8 \\
81 - 88 &= -7
\end{aligned}
$$

b) a média das diferenças

$$\bar{d} = -\frac{30}{9} = -3{,}333$$

c) a variância das diferenças

$$s^2 = \frac{200}{8} = 25$$

d) o valor de t

$$t = \frac{-3{,}333}{\sqrt{\frac{25}{9}}} = -2{,}0$$

que está associado a $n - 1 = 9 - 1 = 8$ graus de liberdade.

Ao nível de significância de 1% e com 8 graus de liberdade, o valor de t na Tabela A.6 do Apêndice é 3,36. Como o valor absoluto de t calculado (2,0) é menor do que o valor da tabela (3,36), conclui-se que o tratamento não tem efeito significante, ao nível de 1%. Em termos práticos, o experimento não provou que a dieta emagrece.

12.5 - Teste t para Observações Independentes Quando as Variâncias São Desiguais

O teste t, na forma apresentada na Seção 12.1, só deve ser aplicado quando as variâncias das populações são iguais. Mas como se estabelece que as variâncias das populações são iguais?

Convém saber que existe uma regra prática: comparam-se as variâncias das duas amostras; se a maior variância for igual até 4 vezes a menor, admite-se que as duas populações têm variâncias iguais. Por exemplo, se as amostras têm variâncias $s_1^2 = 15{,}64$ e $s_2^2 = 6{,}80$, tem-se que

$$\frac{s_1^2}{s_2^2} = \frac{15{,}64}{6{,}80} = 2{,}30 < 4,$$

ou seja, é razoável admitir que as variâncias são iguais. Mas é melhor aplicar um teste estatístico.

Para testar a hipótese de que as variâncias das duas populações são iguais, aplica-se o teste F. Para isso, é preciso, primeiro, estabelecer o nível de significância. Depois, é preciso calcular:

a) a variância de cada grupo, indica-se:

s_1^2 : variância do grupo 1

s_2^2 : variância do grupo 2

b) o valor de F, dado pela razão entre a maior e a menor variância. Se $s_1^2 > s_2^2$, o valor

$$F = \frac{s_1^2}{s_2^2}$$

está associado a $n - 1$ (numerador) e $n_2 - 1$ (denominador) graus de liberdade.

Feitos os cálculos, é preciso procurar o valor de F na tabela, com nível de significância igual à metade do nível de significância estabelecido, e com $(n_1 - 1)$ e $(n_2 - 1)$ graus de liberdade. Toda vez que o valor calculado de F for igual ou maior do que o valor da tabela, rejeita-se a hipótese de que as variâncias das duas populações são iguais, ao nível de significância estabelecido.

Para entender como se acha o valor de F na tabela, observe a Figura 12.2, que reproduz parte dessa tabela, apresentada neste livro, em Apêndice. Foi sombreado o valor de F ao nível de significância de 2,5% e com 7 e 8 graus de liberdade, que seria utilizado para um teste na forma descrita aqui, mas ao nível de significância de 5% e com os mesmos graus de liberdade.

Figura 12.2 *Valor de F para $\alpha = 2{,}5\%$, com 7 e 8 graus de liberdade*

Valores de F para $\alpha = 2{,}5\%$, segundo o número de graus de liberdade do numerador e do denominador

Nº de g. l. do denominador	Número de graus de liberdade do numerador								
	1	2	3	4	5	6	7	8	9
1	648,0	800,0	864,0	900,0	922,0	937,0	948,0	957,0	963,0
2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4
3	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,5
4	12,2	10,6	9,98	9,60	9,36	9,20	9,07	8,98	8,90
5	10,0	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03

Se as variâncias são desiguais, para comparar duas médias aplica-se o teste t, na forma descrita aqui. É preciso calcular:

a) a média de cada grupo. Indica-se

$\bar{x}_1$: média do grupo 1
$\bar{x}_2$: média do grupo 2

b) a variância de cada grupo, indica-se

s_1^2 : variância do grupo 1

s_2^2 : variância do grupo 2

c) o valor de t, dado pela fórmula:

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

onde n_1 é o número de elementos do grupo 1 e n_2 é o número de elementos do grupo 2.

d) o número de graus de liberdade associado ao valor de t, que é a parte inteira do número g, obtido pela fórmula:

$$g = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Feitos os cálculos, é preciso procurar o valor de t na tabela, ao nível de significância estabelecido e com g graus de liberdade. Toda vez que o valor absoluto de t calculado for igual ou maior do que o valor na tabela, conclui-se que as médias não são iguais, ao nível de significância estabelecido.

12.6 - Exemplo de Aplicação

Para verificar se determinada dieta leva à perda de peso um médico separou, ao acaso, um conjunto de pacientes em dois grupos: um grupo foi submetido à dieta (grupo tratado), enquanto o outro manteve os mesmos hábitos alimentares (grupo controle). Decorrido determinado período de tempo, o médico obteve a perda de peso de cada paciente, em cada grupo. Os valores estão na Tabela 12.3.

Tabela 12.3

Perdas de peso em quilogramas de pacientes segundo o grupo

Grupo	
Tratado	Controle
12	1
14	0
12	0
9	1
14	0,5
14	1
9	0

Para proceder ao teste, é preciso, primeiro, estabelecer o nível de significância. Seja $\alpha = 5\%$. Depois é preciso calcular:

a) a média de cada grupo:

$$\bar{x}_1 = \frac{12 + 14 + \ldots + 9}{7} = 12$$

$$\bar{x}_2 = \frac{1 + 0 + \ldots + 0}{7} = 0{,}5$$

b) a variância de cada grupo:

$$s_1^2 = \frac{1038 - \frac{(84)^2}{7}}{6} = 5{,}00$$

$$s_2^2 = \frac{3{,}25 - \frac{(3{,}5)^2}{7}}{6} = 0{,}25$$

c) o valor de F, porque como as variâncias são muito diferentes, convém fazer o teste. Seja $\alpha = 5\%$.

$$F = \frac{s_1^2}{s_2^2} = \frac{5}{0{,}25} = 20{,}00$$

O valor calculado de F está associado a 6 (numerador) e 6 (denominador) graus de liberdade. Na Tabela A.3 do Apêndice, de F para $\alpha = 2{,}5\%$, com 6 e 6 graus de liberdade, encontra-se o valor 5,82. Então se rejeita a hipótese de que as variâncias são iguais ao nível de significância de 5%. Agora é preciso calcular:

d) o valor de t:

$$t = \frac{0{,}5 - 12}{\sqrt{\frac{5{,}0}{7} + \frac{0{,}25}{7}}}$$

$$t = \frac{-11{,}5}{\sqrt{\frac{5{,}25}{7}}} = -13{,}28$$

e) o número de graus de liberdade:

$$g = \frac{\left(\frac{5{,}0}{7} + \frac{0{,}25}{7}\right)^2}{\frac{\left(\frac{5{,}0}{7}\right)^2}{6} + \frac{\left(\frac{0{,}25}{7}\right)^2}{6}}$$

$$= \frac{0{,}5625}{0{,}085247} = 6{,}6$$

O valor calculado de t está associado a aproximadamente 6 graus de liberdade. Como o valor de t na Tabela A.6 do Apêndice, ao nível de significância de 5% e com 6 graus de liberdade, é 2,45, rejeita-se a hipótese de que as médias são iguais. Em termos práticos, a perda de peso foi, em média, significativamente maior no grupo submetido à dieta.

12.7 - Teste t para o Coeficiente de Correlação

O teste t, apresentado neste Capítulo, tem outros usos, além da comparação de médias. Por exemplo, o teste t pode ser usado para testar a hipótese de que o coeficiente de correlação entre duas variáveis é igual a zero, contra a hipótese de que é diferente de zero.

Reveja a Seção 6.3 do Capítulo 6. O coeficiente de correlação varia entre -1 e $+1$. Se o coeficiente de correlação entre duas variáveis for igual a zero, não existe correlação linear entre elas. Mas e se o coeficiente calculado for $r = 0{,}30$? Não se pode julgar o valor desse coeficiente sem saber o tamanho da amostra. Quando a amostra é muito pequena, mesmo coeficientes de correlação com valores altos têm pouco significado.

É claro que, se o coeficiente de correlação entre duas variáveis for igual a zero, não existe correlação linear entre elas. Mas se o coeficiente calculado for $r = -0{,}775$ (veja o exercício 6.4.1). Para aplicar o teste t, usa-se a fórmula

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

onde r é o valor calculado para o coeficiente de correlação e n é o tamanho da amostra. Esse valor de t está associado a $n - 2$ graus de liberdade. No caso do exemplo:

$r = -0{,}775$
$n = 14$
Então:

$$t = \frac{-0{,}775}{\sqrt{1-0{,}601}}\sqrt{14-2} = \frac{-0{,}775}{0{,}632}\cdot 3{,}46 = -4{,}25$$

com $n - 2 = 12$ graus de liberdade

Ao nível de significância de 5% a Tabela A.6 do Apêndice dá, para 12 graus de liberdade, o valor $t = 2{,}18$. Como o valor calculado de t é, em valor absoluto, maior do que 2,18, a correlação entre as variáveis é significante ao nível de 5%.

12.8.1 - *Os valores apresentados na Tabela 12.4 permitem testar a hipótese de que recém-nascidos de ambos os sexos têm, em média, a mesma estatura. Teste essa hipótese, ao nível de significância de 5%.*

Tabela 12.4

Tamanho da amostra, média e variância da estatura, em centímetros, de recém-nascidos, segundo o sexo

Sexo	n	$\bar{x}$	s^2
Masculino	1 442	49,29	5,76
Feminino	1 361	48,54	6,30

Fonte: ARENA (1976)

Antes de proceder ao teste t, convém testar a igualdade das variâncias. Para isso, calcula-se:

$$F = \frac{6{,}30}{5{,}76} = 1{,}09$$

que está associado a 1 360 (numerador) e 1 441 (denominador) graus de liberdade. Para proceder ao teste F, ao nível de significância de 5% é preciso procurar, na tabela de F com $\alpha = 2{,}5\%$, o valor associado a 1 360 e 1 441 graus de liberdade. Mas a tabela não tem esses números de graus de liberdade. Como os números são muito grandes, usa-se o valor de F associado a infinitos graus de liberdade, tanto para numerador como para denominador. Esse valor é 1,00. O valor calculado de F é maior do que 1,00. Rejeita-se a hipótese de que as variâncias são iguais, ao nível de significância de 5%.

O teste t — no caso de variâncias desiguais — deve ser calculado como segue:

$$t = \frac{49{,}29 - 48{,}54}{\sqrt{\frac{5{,}76}{1442} + \frac{6{,}30}{1361}}} = 8{,}076$$

que está associado aos graus de liberdade

$$g = \frac{\left(\frac{5{,}76}{1442} + \frac{6{,}30}{1361}\right)^2}{\frac{\left(\frac{5{,}76}{1442}\right)^2}{1441} + \frac{\left(\frac{6{,}30}{1361}\right)^2}{1360}} = 2772$$

Como o valor calculado de t é maior do que o valor dado na Tabela A.6 do Apêndice, rejeita-se, ao nível de significância de 5%, a hipótese de que recém-nascidos de ambos os sexos têm, em média, a mesma estatura. Em termos práticos, os meninos nascem com estatura maior do que as meninas.

12.8.2 - ***Com base nos dados apresentados na Tabela 12.5 teste, ao nível de significância de 5%, a hipótese de que o calibre da veia esplênica é, em média, o mesmo, antes e após a oclusão da veia porta.***

Note que foram tomadas duas medidas em cada cão: uma antes, outra após a oclusão da veia porta. Para aplicar o teste t é preciso calcular a diferença observada em cada animal. Tais diferenças estão na Tabela 12.6.

Tabela 12.5
Calibre da veia esplênica em seis cães antes e após a oclusão da veia porta

Número do cão	Oclusão da veia porta	
	Antes	Depois
1	75	85
2	50	75
3	50	70
4	60	65
5	50	60
6	70	90

Fonte: HOSSNE (1958)

Tabela 12.6
Diferenças de calibre da veia esplênica antes e após a oclusão da veia porta

Número do cão	Oclusão da veia porta		
	Antes	Depois	Diferença
1	75	85	10
2	50	75	25
3	50	70	20
4	60	65	5
5	50	60	10
6	70	90	20

A média das diferenças é:

$$\overline{d} = 15{,}0$$

e a variância é:

$$s^2 = 60{,}00.$$

O valor de t, associado a 5 graus de liberdade, é:

$$t = \frac{15{,}0}{\sqrt{\frac{60{,}00}{6}}} = 4{,}74$$

Na tabela de t, para $\alpha = 5\%$ e com 5 graus de liberdade, encontra-se o valor 2,57. Como o valor calculado de t é maior do que o da tabela, rejeita-se, ao nível de significância de 5%, a hipótese de que o calibre da veia esplênica é, em média, o mesmo, antes e depois da oclusão da veia porta. Em termos práticos, a oclusão da veia porta determina aumento significativo do calibre da veia esplênica.

12.9 - Exercícios Propostos

12.9.1 *Dez ratos machos adultos, criados em laboratório, foram separados aleatoriamente em dois grupos: um grupo foi tratado com a ração normalmente usada no laboratório e o outro grupo foi submetido a uma nova ração (experimental). Decorrido certo período de tempo, pesaram-se os ratos. Os pesos estão apresentados na Tabela 12.7. Teste a hipótese de que o peso médio dos ratos é o mesmo, para os dois tipos de ração.*

Tabela 12.7
Pesos em gramas de ratos adultos, segundo a ração

Ração	
Padrão	Experimental
200	220
180	200
190	210
190	220
180	210

12.9.2 - *Os quocientes de inteligência (QI) de 10 crianças, segundo dois testes de inteligência, A e B, estão apresentados na Tabela 12.8. Verifique, através do teste t, se os dois testes de inteligência dão, em média, o mesmo valor.*

12.9.3 - *A Tabela 12.9 apresenta dados de pressão sangüínea sistólica de mulheres na faixa etária de 30 a 35 anos, que usavam e não usavam anovulatório. Teste a hipótese de que o uso de anovulatórios não tem efeito sobre a pressão sangüínea sistólica.*

12.9.4 - *A Tabela 12.10 apresenta o tamanho da amostra, a média e a variância dos pesos ao nascer de nascidos vivos de ambos os sexos. Teste, ao nível de significância de 1%, a hipótese de que os dois sexos têm, em média, o mesmo peso ao nascer.*

Tabela 12.8

Valores de QI em dez crianças, segundo o teste de inteligência aplicado

Teste	
A	B
100	105
105	108
98	102
101	103
100	100
108	110
98	106
100	100
99	103
99	103

Tabela 12.9

Pressão sangüínea sistólica de mulheres de 30 a 35 anos segundo o uso de anovulatório

Uso de anovulatório	
Sim	Não
111	109
119	113
121	120
113	117
116	108
126	120
128	122
123	124
122	115
121	112

Tabela 12.10

Tamanho da amostra, média e variância de pesos ao nascer de nascidos vivos segundo o sexo

Sexo	n	$\overline{x}$	s^2
Masculino	14	3,253	0,261
Feminino	13	3,130	0,265

CAPÍTULO
13

Análise de Variância

O Capítulo 12 explica como comparar médias de duas populações, com base em amostras dessas populações. Mas às vezes é preciso comparar médias de *mais de duas populações*. Por exemplo, para verificar se pessoas com diferentes níveis de renda, isto é, alto, médio e baixo têm, em média, o mesmo peso corporal, é preciso comparar médias de três populações.

Outras vezes, é preciso comparar várias situações experimentais. Por exemplo, se um pesquisador separa, ao acaso, um conjunto de pacientes em 4 grupos e administra uma droga diferente a cada grupo, terá que comparar médias de quatro "populações".

Para comparar médias de mais de duas populações aplica-se o teste F, na forma descrita nesse Capítulo, desde que a variável em estudo tenha distribuição normal ou aproximadamente normal. Mas antes de mostrar como se faz esse teste, convém apresentar um exemplo.

Imagine que 4 amostras casuais simples, todas com cinco elementos mas cada uma proveniente de uma população, conduziram aos dados apresentados na Tabela 13.1. As médias dessas amostras estão na última linha dessa tabela. Será que as diferenças das médias das amostras são suficientemente grandes para que se possa afirmar que as médias das populações são diferentes? Para responder a esta pergunta, é preciso um teste estatístico.

Tabela 13.1
Dados de 4 amostras e respectivas médias

Amostras			
A	B	C	D
11	8	5	4
8	5	7	4
5	2	3	2
8	5	3	0
8	5	7	0
8	5	5	2

13.1 - Análise de Variância para Experimentos ao Acaso

Se a variável em estudo tem distribuição normal ou aproximadamente normal, para comparar mais de duas médias aplica-se o teste F. Primeiro, é preciso estudar as *causas de variação*. Por que os dados variam? Uma explicação é o fato de as amostras provirem de populações diferentes. Outra explicação é o acaso, porque mesmo dados provenientes da mesma população variam.

O teste F é feito através de uma *análise de variância*, que separa a variabilidade devido aos "tratamentos" (no exemplo, devido às amostras terem provindo de populações diferentes) da variabilidade residual, isto é, devido ao acaso. Para aplicar o teste F é preciso fazer uma série de cálculos, que exigem conhecimento da notação.

A Tabela 13.2 apresenta os dados de k tratamentos, cada um com r repetições (no exemplo, denominam-se repetições os elementos da mesma amostra). A soma das r repetições de um mesmo tratamento constitui o *total* desse tratamento. O *total geral* é dado pela soma dos k totais de tratamentos.

Para fazer a análise de variância é preciso calcular as seguintes quantidades:

a) os graus de liberdade:
de tratamentos: $k - 1$
do total: $n - 1$
de resíduo: $(n - 1) - (k - 1) = n - k$

b) o valor C, dado pelo total geral elevado ao quadrado e dividido pelo número de dados. O valor C é chamado *correção*.

$$C = \frac{(\sum x)^2}{n}$$

c) a soma de quadrados total:

$$SQT = \sum x^2 - C$$

d) a soma de quadrados de tratamentos:

$$SQTr = \frac{\sum T^2}{r} - C$$

e) a soma de quadrados de resíduo:

$$SQR = SQT - SQTr$$

f) o quadrado médio de tratamentos:

$$QMTr = \frac{SQTr}{k-1}$$

g) o quadrado médio de resíduo:

$$QMR = \frac{SQR}{n-k}$$

h) o valor de F

$$F = \frac{QMTr}{QMR}$$

Tabela 13.2
Notação para a análise de variância

	Tratamento					
	1	*2*	*3*	...	*k*	Total
	x_{11}	x_{21}	x_{31}		x_{k1}	
	x_{12}	x_{22}	x_{32}		x_{k2}	
	.	.	.		.	
	.	.	.		.	
	.	.	.		.	
	x_{1r}	x_{2r}	x_{3r}		x_{kr}	
Total	T_1	T_2	T_3		T_k	$\Sigma T = \Sigma x$
Nº de repetições	r	r	r		r	$n = kr$
Média	$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$		$\overline{x}_k$	

Em seguida, é preciso comparar o valor calculado de F com o valor dado na tabela de F, ao nível de significância estabelecido e com $(k-1)$ graus de liberdade no numerador e $(n-k)$ graus de liberdade no denominador. Toda vez que o valor calculado de F for igual ou maior do que o dado na tabela de F conclui-se, ao nível de significância estabelecido, que as médias de tratamentos não são iguais.

Para entender como se procura o valor de F na tabela, observe a Figura 13.1, que reproduz parte da Tabela A.4 dada neste livro, em Apêndice. Está sombreado o valor de F ao nível de significância de 5%, com 3 graus de liberdade para tratamentos (numerador) e 8 graus de liberdade para resíduo (denominador).

Figura 13.1 *Valor de F ao nível de significância de 5%, com 3 e 8 graus de liberdade*

	1	2	3	4
1	161	200	216	225
2	18,5	19,0	19,2	19,2
3	10,1	9,55	9,28	9,12
4	7,71	6,94	6,59	6,39
5	6,61	5,79	5,41	5,19
6	5,99	5,14	4,76	4,53
7	5,59	4,74	4,35	4,12
8	5,32	4,46	4,07	3,84
9	5,12	4,26	3,86	3,63

Um exemplo ajuda a entender como se aplica o teste F para a comparação de médias. Veja os dados apresentados na Tabela 13.1. Para fazer a análise de variância é preciso calcular:

a) os graus de liberdade:
de tratamentos: $k-1=4-1=3$
do total: $n-1=20-1=19$
de resíduo: $n-k=20-4=16$

b) o valor de C:

$$C=\frac{(11+8+\ldots+0)^2}{20}=\frac{100^2}{20}=500$$

c) a soma de quadrados total:

$$SQT = 11^2 + 8^2 + ... + 0^2 - 500$$
$$= 658 - 500$$
$$= 158$$

d) a soma de quadrados de tratamentos:

$$SQTr = \frac{40^2 + 25^2 + 25^2 + 10^2}{5} - 500$$
$$= 590 - 500$$
$$= 90$$

e) a soma de quadrados de resíduo:

$$SQR = 158 - 90$$
$$= 68$$

f) o quadrado médio de tratamentos:

$$QMTr = \frac{90}{3} = 30$$

g) o quadrado médio do resíduo:

$$QMR = \frac{68}{16} = 4{,}25$$

h) o valor de F:

$$F = \frac{30}{4{,}25} = 7{,}06$$

As quantidades calculadas são apresentadas numa *tabela de análise de variância*. Veja a Tabela 13.3.

Tabela 13.3
Análise de variância dos dados da Tabela 13.1

Causas de variação	GL	SQ	QM	F
Tratamentos	3	90	30	7,06
Resíduo	16	68	4,25	
Total	19	158		

Ao nível de significância de 5%, o valor de F na Tabela A.4 do Apêndice, com 3 e 16 graus de liberdade, é 3,24. Como o valor obtido é maior do que 3,24, conclui-se que as médias não são iguais, ao nível de significância de 5%.

13.2 - Teste de Tukey para Comparação de Médias

Uma análise de variância permite estabelecer se as médias das populações em estudo são, ou não são, estatisticamente iguais. No entanto, esse tipo de análise não permite detectar quais são as médias estatisticamente diferentes das demais. Por exemplo, a análise de variância apresentada na Tabela 13.3 mostrou que as médias das populações não são iguais, mas não permite concluir qual é, ou quais são, as médias diferentes das demais.

O teste de Tukey permite estabelecer a *diferença mínima significante*, ou seja, a menor diferença de médias de amostras que deve ser tomada como estatisticamente significante, em determinado nível. Essa diferença (d.m.s.) é dada por:

$$\text{d. m. s.} = q\sqrt{\frac{\text{QMR}}{r}}$$

onde q é um valor dado em tabela, QMR é o quadrado médio do resíduo da análise de variância e r é o número de repetições de cada tratamento.

A tabela de q é dada neste livro, em Apêndice. Para entender como se usa essa tabela, primeiro observe a Figura 13.2. O valor sombreado seria utilizado para comparar as médias de 3 tratamentos, quando o número de graus de liberdade do resíduo da análise de variância é igual a 6, ao nível de significância de 5%.

Figura 13.2 *Valor de q para α = 5%, 3 tratamentos e 6 graus de liberdade no resíduo*

Nº de graus de lib. do resíduo	2	3	4	5
1	18,0	27,0	32,8	37,1
2	6,08	8,33	9,80	10,9
3	4,50	5,91	6,82	7,50
4	3,93	5,04	5,76	6,29
5	3,64	4,60	5,22	5,67
6	3,46	4,34	4,90	5,30
7	3,34	4,16	4,68	5,06

Considere agora os dados da Tabela 13.1. A análise de variância apresentada na Tabela 13.3 mostra um valor de F significante ao nível de 5%. Então as médias de A, B, C e D não são estatisticamente iguais. Mas qual é, ou quais são, as médias diferentes entre si?

A pergunta pode ser respondida com a aplicação do teste de Tukey. Ao nível de significância de 5%, o valor de q para comparar 4 tratamentos (A, B, C e D), com 16 graus de liberdade no resíduo (veja a Tabela 13.3), é 4,05. Como QMR = 4,25 e $r = 5$, segue-se que:

$$\text{d.m.s.} = 4{,}05 \sqrt{\frac{4{,}25}{5}}$$

$$= 3{,}73$$

De acordo com o teste de Tukey, duas médias são estatisticamente diferentes toda vez que o valor absoluto da diferença entre elas for igual ou superior ao valor da d.m.s. No caso do exemplo, o valor da d.m.s. é 3,73 e os valores absolutos das diferenças entre as médias estão apresentados na Tabela 13.4. É fácil ver que a diferença entre as médias A e D é maior do que a d.m.s. Então, ao nível de 5%, a média de A é significativamente maior do que a média de D.

Tabela 13.4

Valores absolutos das diferenças entre as médias dos tratamentos A, B, C e D

Pares de médias	Valor absoluto da diferença
A e B	8 – 5 = 3
A e C	8 – 5 = 3
A e D	8 – 2 = 6
B e C	5 – 5 = 0
B e D	5 – 2 = 3
C e D	5 – 2 = 3

13.3 - Análise de Variância com Número Diferente de Repetições

Muitas vezes o pesquisador dispõe de diversas amostras, cada uma proveniente de uma população, mas essas amostras não têm todas o mesmo tamanho. Mesmo assim, é possível conduzir a análise de variância. Aliás, todos os cálculos, com exceção da soma de quadrados de tratamentos, são feitos na forma já apresentada na Seção 13.2.

Para entender como se calcula a soma de quadrados de tratamentos — quando os tratamentos não têm o mesmo número de repetições — primeiro observe a Tabela 13.5.

Tabela 13.5
Notação para a análise de variância

	Tratamento					
	1	2	3	...	k	Total
	x_{11}	x_{21}	x_{31}		x_{k1}	
	x_{12}	x_{22}	x_{32}		x_{k2}	
	.	.	.		.	
	.	.	.		.	
	.	.	.		.	
	x_{1r_1}	x_{2r_2}	x_{3r_3}		x_{kr_k}	
Total	T_1	T_2	T_3		T_k	$\Sigma T = \Sigma x$
Nº de repetições	r_1	r_2	r_3		r_k	$n = \Sigma r$
Média	$\overline{x_1}$	$\overline{x_2}$	$\overline{x_3}$		$\overline{x_k}$	

A soma de quadrados de tratamentos é dada pela fórmula:

$$\text{SQTr} = \frac{T_1^2}{r_1} + \frac{T_2^2}{r_2} + \ldots + \frac{T_k^2}{r_k} - C,$$

onde C é a correção, definida na Seção 13.1.

É mais fácil entender a aplicação de fórmulas através de um exemplo. Veja então os dados apresentados na Tabela 13.6. Note que o tratamento A tem 6 repetições, o tratamento B tem 4 repetições e o tratamento C tem 5 repetições.

Para fazer a análise de variância dos dados apresentados na Tabela 13.6, é preciso calcular:

a) os graus de liberdade:
   do total: $15 - 1 = 14$
   de tratamentos: $3 - 1 = 2$
   do resíduo: $14 - 2 = 12$

b) o valor de C:

$$C = \frac{245^2}{15} = 4001{,}67$$

Tabela 13.6

Dados, segundo o tratamento e os respectivos totais

Tratamento		
A	B	C
15	23	19
10	16	15
13	19	21
18	18	14
15		16
13		
84	76	85

c) a soma de quadrados total:

$$SQT = 15^2 + 10^2 + ... + 16^2 - 4\,001{,}67 = 159{,}33$$

d) a soma de quadrados de tratamentos:

$$SQTrat = \frac{84^2}{6} + \frac{76^2}{4} + \frac{85^2}{5} - 4001{,}67 = 63{,}33$$

e) a soma de quadrados de resíduo:

$$SQR = SQT - SQTr$$
$$= 159{,}33 - 63{,}33 = 96{,}00$$

f) o quadrado médio de tratamentos:

$$QMTr = \frac{63{,}33}{2} = 31{,}67$$

g) o quadrado médio de resíduo:

$$QMR = \frac{96{,}00}{12} = 8{,}00$$

h) o valor de F:

$$F = \frac{31{,}67}{8{,}00} = 3{,}96$$

Os valores calculados estão apresentados na Tabela 13.7

Tabela 13.7

Análise de variância dos dados apresentados na Tabela 13.6

Causas de variação	GL	SQ	QM	F
Tratamentos	2	63,33	31,67	3,96
Resíduo	12	96,00	8,00	
Total	14	159,33		

Ao nível de significância de 5%, com dois e 12 graus de liberdade, o valor de F dado na Tabela A.4 do Apêndice é 3,89. Como o valor calculado de F é 3,96, maior do que 3,89, conclui-se que as médias não são iguais.

Para comparar as médias de tratamentos duas a duas, aplica-se o teste de Tukey que, neste caso, é aproximado, porque os tratamentos têm número diferente de repetições. A diferença mínima significativa (d.m.s.) é dada pela fórmula:

$$\text{d. m. s.} = q\sqrt{\left(\frac{1}{r_i}+\frac{1}{r_j}\right)\frac{\text{QMR}}{2}}$$

onde r_i é o número de repetições do i-ésimo tratamento e r_j é o número de repetições do j-ésimo tratamento.

No caso do exemplo, para comparar a média de A com a média de B, tem-se:

$$\text{d. m. s.} = 3{,}77\sqrt{\left(\frac{1}{6}+\frac{1}{4}\right)\frac{8{,}00}{2}}$$

$$= 4{,}87$$

Para comparar A com C, tem-se:

$$\text{d. m. s.} = 3{,}77\sqrt{\left(\frac{1}{6}+\frac{1}{5}\right)\frac{8{,}00}{2}}$$

$$= 4{,}57$$

Para comparar B com C, tem-se:

$$\text{d. m. s.} = 3{,}77\sqrt{\left(\frac{1}{4}+\frac{1}{5}\right)\frac{8{,}00}{2}}$$

$$= 5{,}06$$

Os valores absolutos das diferenças entre as médias estão na Tabela 13.8. Como o valor absoluto da diferença entre A e B é maior do que a respectiva d.m.s., conclui-se que, em média, A é diferente de B, ao nível de significância de 5%.

Tabela 13.8

Valores das diferenças entre as médias

Pares de médias	Valor absoluto da diferença
A e B	$\|14 - 19\| = 5$
A e C	$\|14 - 17\| = 3$
B e C	$\|19 - 17\| = 2$

13.4 - Exercícios Resolvidos

13.4.1 - *Com base nos dados apresentados na Tabela 13.9, verifique se existe diferença estatística entre os grupos. Note que são três grupos em comparação. No grupo operado foi feita a remoção das glândulas salivares maiores, e no grupo pseudo-operado foram executados todos os tempos cirúrgicos, mas nenhuma glândula foi removida.*

Tabela 13.9

Taxa de glicose, em miligramas por 100ml de sangue, em ratos Wistar machos de 60 dias, segundo o grupo

Grupo		
Operado	Pseudo-operado	Normal
96,0	90,0	86,0
95,0	93,0	85,0
100,0	89,0	105,0
108,0	88,0	105,0
120,0	87,0	90,0
110,5	92,5	100,0
97,0	87,5	95,0
92,5	85,0	95,0

Fonte: GUIMARÃES et alii (1979)

Para fazer a análise de variância, é preciso calcular:

$$C = \frac{2292{,}0^2}{24} = 218\,886{,}00$$

$$SQT = 220\,722{,}00 - 218\,886{,}00 = 1\,836{,}00$$

$$SQTr = \frac{1756826{,}00}{8} - 218\,886{,}00 = 717{,}25$$

$$SQR = 1\,836{,}00 - 717{,}25 = 1\,118{,}75$$

$$QMT = \frac{717,25}{2} = 358,625$$

$$QMR = \frac{1118,75}{21} = 53,274$$

$$F = \frac{358,625}{53,274} = 6,73$$

Estes valores estão apresentados na Tabela 13.10. O valor de F é significante ao nível de 5%.

Tabela 13.10
Análise de variância dos dados da Tabela 13.9

Causas de variação	GL	SQ	QM	F
Grupos	2	717,25	358,625	6,73
Resíduo	21	1 118,75	53,274	
Total	23	1 836,00		

Para aplicar o teste de Tukey ao nível de significância de 5%, é preciso obter o valor de q, para 3 tratamentos e 21 graus de liberdade no resíduo. Como na tabela não existe esse valor, faz-se uma interpolação. Com 3 tratamentos e 20 graus de liberdade no resíduo, $q = 3,58$; com 3 tratamentos e 24 graus de liberdade do resíduo, $q = 3,53$. Então, aumentando $24 - 20 = 4$ graus de liberdade, o valor de q diminuiu $3,58 - 3,53 = 0,05$. Logo, aumentando de $21 - 20 = 1$ grau de liberdade, o valor q diminuirá x, tal que:

$$4 \rightarrow 0,05$$
$$1 \rightarrow x$$

$$x = \frac{0,05}{4} = 0,0125$$

Portanto, com 3 tratamentos e 21 graus de liberdade no resíduo,

$$q = 3,58 - 0,01 = 3,57$$

Para proceder ao teste de Tukey, é preciso calcular:

$$\text{d.m.s.} = 3,57\sqrt{\frac{53,247}{8}} = 9,21.$$

Os valores absolutos das diferenças de médias estão apresentados na Tabela 13.11. A taxa de glicose é, em média, maior nos operados do que nos pseudo-operados, ao nível de significância de 5%.

Tabela 13.11
Valores absolutos das diferenças de médias

Pares de médias	Diferença (Valor absoluto)
Operado vs pseudo	|102,375 – 89,000| = 13,375
Operado vs normal	|102,375 – 95,125| = 7,250
Pseudo vs normal	| 89,000 – 95,125| = 6,125

13.4.2 - *Faça a análise de variância e aplique o teste de Tukey aos dados apresentados na Tabela 13.12.*

Tabela 13.12
Dados segundo a amostra

Amostra		
A	B	C
24	9	20
19	16	25
26	14	18
25	9	19
22	12	18
26		
23		
27		

Para fazer a análise de variância é preciso calcular:

$$C = \frac{352^2}{18} = 6883{,}56$$

$$\text{SQT} = 7\,448 - 6883{,}56 = 564{,}44$$

$$\text{SQTr} = 7\,328 - 6883{,}56 = 444{,}44$$

$$\text{SQR} = 564{,}44 - 444{,}44 = 120{,}00$$

$$\text{QMTr} = \frac{444{,}44}{2} = 222{,}22$$

$$\text{QMR} = \frac{120{,}00}{15} = 8{,}00$$

$$F = \frac{222{,}22}{8{,}00} = 27{,}78$$

Os valores calculados estão apresentados na Tabela 13.13. O valor de F é significante ao nível de 5%.

Tabela 13.13

Análise de variância dos dados da Tabela 13.12

Causas de variação	GL	SQ	QM	F
Amostras	2	444,44	222,22	27,78
Resíduo	15	120,00	8,00	
Total	17	564,44		

A comparação de médias, 2 a 2, é feita pelo teste de Tukey. Para comparar as médias de A e B, ou as médias de A e C, calcula-se:

$$\text{d.m.s.} = 3{,}67\sqrt{\left(\frac{1}{8}+\frac{1}{5}\right)\frac{8{,}00}{2}} = 4{,}18$$

e, para comparar as médias de B e C, calcula-se:

$$\text{d.m.s.} = 3{,}67\sqrt{\frac{8{,}00}{2}} = 4{,}64$$

Os valores absolutos das diferenças entre as médias das amostras estão na Tabela 13.14. Ao nível de 5%, a média de B é significativamente menor do que as médias de A e C.

Tabela 13.14

Valores absolutos das diferenças entre as médias das amostras

Pares de médias	Valor absoluto da diferença
A e B	$\lvert 24-12 \rvert = 12$
A e C	$\lvert 24-20 \rvert = 4$
B e C	$\lvert 12-20 \rvert = 8$

13.5 — Exercícios Propostos

13.5.1 - *Faça a análise de variância e aplique o teste de Tukey aos dados apresentados na Tabela 13.15.*

13.5.2 - *Faça a análise de variância e aplique o teste de Tukey aos dados apresentados na Tabela 13.16.*

Tabela 13.15
Dados segundo o tratamento

Tratamento				
A	B	C	D	E
17	15	15	15	20
19	8	23	9	13
13	9	17	10	14
19	12	21	14	17

Tabela 13.16
Dados segundo o tratamento

Tratamento			
A	B	C	D
11	1	21	14
20	9	13	5
18	8	11	7
15	2		10

13.5.3 - *Faça a análise de variância dos dados apresentados na Tabela 12.8 do Capítulo 12. Verifique que o valor calculado de F é igual ao quadrado do valor de t, calculado para os mesmos dados. Isto não é coincidência; com 1 grau de liberdade no numerador,* $F = t^2$.

13.5.4 - *Verifique que os valores de F com um grau de liberdade no numerador, apresentados nas tabelas do Apêndice, são iguais aos quadrados dos valores de t, apresentados na tabela do Apêndice, para o mesmo* α *e com os mesmos graus de liberdade. Então, para comparar dois tratamentos, é indiferente usar teste t ou o teste F. Ambos levam à mesma conclusão.*

CAPÍTULO
14

Intervalo de Confiança

Imagine uma amostra casual simples de n elementos. A média dos dados dessa amostra constitui uma *estimativa* da média da população de onde essa amostra proveio. Para indicar a *precisão* dessa estimativa, calcula-se o intervalo de confiança para a média na forma dada neste Capítulo. Mas, antes de conceituar intervalo de confiança, é preciso entender o que é erro padrão da média.

14.1 - Erro Padrão da Média

Uma população é constituída pelos valores 4,10 e 16. A média dessa população, que se indica por μ (lê-se mi), é:

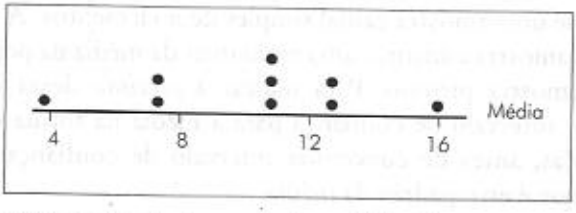
$$\mu = \frac{4+10+16}{3} = 10$$

Considere todas as amostras possíveis de dois elementos que podem ser retirados dessa população, admitindo que todo elemento retirado para compor a amostra é reposto, antes da retirada do segundo. Essas amostras, e as respectivas médias, estão na Tabela 14.1. É fácil ver, observando a Figura 14.1, que as médias das amostras distribuem-se em torno da média $\mu = 10$ da população.

Tabela 14.1
Médias das amostras de dois elementos obtidos da população constituída pelos números 4, 10 e 16

Amostra	Média
4 e 4	4
4 e 10	7
4 e 16	10
10 e 4	7
10 e 10	10
10 e 16	13
16 e 4	10
16 e 10	13
16 e 16	16

Figura 14.1 *Distribuição das médias das amostras*



Para medir o grau de dispersão das médias das amostras em torno da média da população, calcula-se a *variância da média*. Essa medida, que se indica por $\sigma_{\bar{x}}^2$, é dada pela fórmula:

$$\sigma_{\bar{x}}^2 = \frac{\sum_{i=1}^{r} (\bar{x}_i - \mu)^2}{r}$$

onde $\bar{x}_i$ é a média da i-ésima amostra e r o número de amostras que podem ser obtidas da população.

Para as médias apresentadas na Tabela 14.1, a variância da média é:

$$\sigma_{\bar{x}}^2 = \frac{(4-10)^2 + (7-10)^2 + \ldots + (16-10)^2}{9} = \frac{108}{9} = 12$$

Na prática, é impossível calcular a variância da média pela fórmula apresentada, porque o pesquisador dispõe de uma única amostra para estimar a média μ da população e obter uma medida de precisão dessa estimativa — e não de todas as amostras possíveis.

Mas existe uma saída. Já se demonstrou que uma estimativa da variância da média é dada pela fórmula:

$$s_{\bar{x}}^2 = \frac{s^2}{n}$$

onde s^2 é a variância da amostra.

As médias, as variâncias e as variâncias das médias das amostras apresentadas na Tabela 14.1 estão na Tabela 14.2. Note que a média das médias coincide com a média $\mu = 10$ da população e que a média das variâncias das médias das amostras é igual a $\sigma_{\bar{x}}^2 = 12$, calculada anteriormente.

Tabela 14.2

Médias, variâncias e variâncias das médias das amostras apresentadas na Tabela 14.1

Amostra	Média	Variância	Variância da média
4 e 4	4	0	0
4 e 10	7	18	9
4 e 16	10	72	36
10 e 4	7	18	9
10 e 10	10	0	0
10 e 16	13	18	9
16 e 4	10	72	36
16 e 10	13	18	9
16 e 16	16	0	0
Média	10	24	12

Por definição, *erro padrão da média* é a raiz quadrada com sinal positivo da variância da média. Indica-se a estimativa do erro padrão da média por $s_{\bar{x}}$. Logo

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

14.2 - Intervalo de Confiança

Seja X uma variável aleatória com distribuição normal de média μ e variância σ^2. Com base em uma amostra casual simples de n elementos dessa população, obtêm-se as estimativas $\bar{x}$ e s^2, de μ e σ^2, respectivamente.

A expressão

$$\bar{x} - t\frac{s}{\sqrt{n}} < \mu < \bar{x} + t\frac{s}{\sqrt{n}}$$

é o *intervalo de confiança* para a média μ da população. Nessa expressão, t é um valor encontrado na tabela de t (dada neste livro em Apêndice), com $n - 1$ graus de liberdade e ao nível de significância α.

Para melhor entender como se calcula um intervalo de confiança, convém discutir um exemplo. Seja X a variável aleatória que representa a taxa de colesterol no plasma sangüíneo humano. Imagine que, com base em uma amostra casual simples de $n = 25$ indivíduos, foram obtidos a média $\bar{x} = 198\text{mg}/100\text{ml}$ e o desvio padrão $s = 30\text{mg}/100\text{ml}$.

É fácil calcular o intervalo de confiança para μ. Seja $\alpha = 10\%$. O valor de t na Tabela A.6 do Apêndice, com $n - 1 = 25 - 1 = 24$ graus de liberdade, é 1,71. A expressão do intervalo de confiança fica, então, como segue:

$$198 - 1{,}71\frac{30}{\sqrt{25}} < \mu < 198 + 1{,}71\frac{30}{\sqrt{25}}$$

$$187{,}74 < \mu < 208{,}26$$

A interpretação do intervalo de confiança exige cuidado. Quando são obtidas muitas amostras de n elementos de uma mesma população e se determina, para cada amostra, um intervalo de confiança $(100 - \alpha)\%$ desses intervalos contém a média μ da população.

Mas, na prática, o pesquisador dispõe de uma única amostra, que fornece uma estimativa da média e uma estimativa do desvio padrão. Então o pesquisador não sabe se a média da população está, ou não está contida no intervalo que calculou. Sabe, porém, que $(100-\alpha)\%$ dos intervalos de confiança calculados dessa forma contêm a média da população.

14.3 - Alguns Pontos Básicos

O valor $(100 - \alpha)\%$ é denominado *nível de confiança*. Diz-se, por isso, que os intervalos são de confiança $(100 - \alpha)\%$.

Na área biológica, é comum apresentar os valores $\bar{x}$ e $s_{\bar{x}}$ escritos na forma $\bar{x} \pm s_{\bar{x}}$. Esta expressão pode ser vista como um intervalo de confiança para μ, mas com nível de confiança indeterminado. Isto porque o valor de t, nessa expressão, é igual a 1. Mas o valor de t, na tabela, depende do número de graus de liberdade associados a $s_{\bar{x}}$. Então, se n for pequeno, o intervalo $\bar{x} \pm s_{\bar{x}}$ pode ter nível de confiança relativamente baixo. Veja um exemplo.

Considere uma amostra de seis elementos. Para calcular um intervalo de 90% de confiança para μ, usando esta amostra, o valor de t dado na Tabela A.6 do Apêndice é 2,02. Então o intervalo

$$\bar{x} \pm s_{\bar{x}}$$

é menor do que o intervalo de 90% de confiança, que é

$$\bar{x} \pm 2{,}02\, s_{\bar{x}}.$$

14.4.1 - *A taxa de glicose no sangue humano é uma variável aleatória com distribuição aproximadamente normal. Suponha que, com base em uma amostra de 30 pessoas, foi obtida a média $\bar{x} = 102$mg de glicose por 100ml de sangue e o desvio padrão $s = 6$mg. Calcule o intervalo de 90% de confiança para μ.*

O intervalo de confiança é dado pela expressão:

$$\bar{x} - t\frac{s}{\sqrt{n}} < \mu < \bar{x} + t\frac{s}{\sqrt{n}}$$

Na tabela de t, com $30 - 1 = 29$ graus de liberdade e ao nível de significância de $\alpha = 10\%$, encontra-se $t = 1{,}70$. Então o intervalo de confiança fica como segue:

$$102 - 1{,}70\frac{6}{\sqrt{30}} < \mu < 102 + 1{,}70\frac{6}{\sqrt{30}}$$

$$100{,}14 < \mu < 103{,}86$$

14.4.2 - *Com base nos dados apresentados na Tabela 14.3, determine o intervalo de 95% de confiança para μ.*

Tabela 14.3

Taxa de glicose, em miligramas por 100ml de sangue, de ratos Wistar machos de 40 dias

100,0	92,5
87,5	94,0
110,0	100,0
99,5	100,0

Fonte: GUIMARÃES et alii (1979)

A média e o desvio padrão dos dados apresentados na Tabela 14.3 são, respectivamente, $\bar{x} = 97{,}9$ e $s = 6{,}7$. O valor de t, com $8 - 1 = 7$ graus de liberdade e para $\alpha = 5\%$ é 2,36. Então o intervalo de 95% de confiança para μ é:

$$97{,}9 - 2{,}36\frac{6{,}7}{\sqrt{8}} < \mu < 97{,}9 + 2{,}36\frac{6{,}7}{\sqrt{8}}$$

$$92{,}3 < \mu < 103{,}5$$

14.5 - Exercícios Propostos

14.5.1 - *Seja X a variável aleatória que representa a pressão sangüínea sistólica em indivíduos com idade entre 20 e 25 anos. Essa variável tem distribuição aproximadamente normal. Suponha que, com base em uma amostra de 100 indivíduos, foi obtida a média $\bar{x}$ = 123mm de mercúrio e o desvio padrão s = 8 milímetros de mercúrio. Determine o intervalo de 90% de confiança para μ.*

14.5.2 - *Seja X a variável aleatória que representa a taxa de hemoglobina em mulheres. Imagine que, com base em uma amostra aleatória de 20 mulheres, obteve-se a média $\bar{x}$ = 16,2g de hemoglobina por 100ml de sangue e o desvio padrão s = 1,1g. Determine o intervalo de 99% de confiança para μ, supondo que X é uma variável com distribuição normal.*

14.5.3 - *Seja X a variável aleatória que representa a estatura ao nascer para o sexo masculino. Com base em 28 recém-nascidos masculinos, obtiveram-se $\bar{x}$ = 50cm e s = 2,5 cm. Supondo distribuição normal, determine o intervalo de 90% de confiança para μ.*

14.5.4 - *Seja X a variável aleatória que representa a taxa de glicose no sangue humano. Determine o intervalo de 95% de confiança para μ, supondo que uma amostra de 25 pessoas forneceu a média $\bar{x}$ = 95mg de glicose por 100 ml de sangue e o desvio padrão s = 6mg. Suponha que X tem distribuição normal.*

CAPÍTULO
15

Elementos de Matemática

Para ler este livro não é preciso conhecimento de matemática além do que é dado em cursos do segundo grau. Mesmo assim, neste Capítulo são apresentados alguns conceitos usados no texto, e é dada uma noção sobre somatórios.

15.1 - Somatórios

Muitas vezes é preciso indicar a soma de n valores. Como exemplo, considere que n alunos fizeram uma prova e existe interesse em calcular a média das notas obtidas. Deve-se então somar todas as notas e dividir o total por n. Mas a operação de soma pode ser indicada de maneira bastante compacta.

Para entender a indicação, primeiro imagine que os nomes dos alunos estão organizados em uma lista, por ordem alfabética. Faça x_1 indicar a nota do primeiro aluno, x_2 indicar a nota do segundo aluno e assim por diante, até x_n, que irá indicar a nota do n-ésimo aluno. Então a soma das notas dos n alunos poderia ser indicada como segue:

$$x_1 + x_2 + ... + x_n,$$

onde os pontos significam "e assim por diante". Entretanto, a maneira mais compacta de indicar essa soma é como segue:

$$\sum_{i=1}^{n} x_i,$$

que se lê "*somatório de x_i, i de 1 a n*". O símbolo Σ, que indica o somatório, é a letra grega maiúscula sigma. É importante notar que o índice i assume valores inteiros, começando por 1 e terminando por n.

Apenas como exemplo, considere quatro números: $x_1 = 2, x_2 = 4$, $x_3 = 3$ e $x_4 = 1$. É fácil ver que, neste caso:

$$\sum_{i=1}^{4} x_i = 2 + 4 + 3 + 1 = 10$$

Em estatística, muitas vezes é preciso indicar o *quadrado da soma*. Para indicar o quadrado da soma, isto, é para indicar

$$(x_1 + x_2 \ldots + x_n)^2$$

usando o somatório, basta escrever:

$$\left(\sum_{i=1}^{n} x_i \right)^2$$

Então, dados os números $x_1 = 3, x_2 = 4, x_3 = 1, x_4 = 2$ e $x_5 = 3$, é fácil ver que:

$$\left(\sum_{i=1}^{5} x_i \right)^2 = (3 + 4 + 1 + 2 + 3)^2 = 13^2 = 169$$

Outras vezes é preciso indicar uma *soma de quadrados*. Para indicar a soma de quadrados

$$x_1^2 + x_2^2 + \ldots + x_n^2,$$

usando somatório, basta escrever:

$$\sum_{i=1}^{n} x_i^2.$$

Considere os números $x_1 = 3, x_2 = 5$ e $x_3 = 1$. É fácil ver que a soma de quadrados é:

$$\sum_{i=1}^{n} x_i^2 = 3^2 + 5^2 + 1^2 = 35$$

Finalmente, suponha dois conjuntos de números: $x_1, x_2, \ldots x_n$ e y_1, y_2, $\ldots y_n$. Pode haver interesse em obter a *soma dos produtos*

$$x_1 y_1 + x_2 y_2 + \ldots x_n y_n.$$

Usando somatório, escreve-se:

$$\sum_{i=1}^{n} x_i y_i.$$

Como exemplo, considere os números $x_1 = 2, x_2 = 3$ e $x_3 = 0$ e os números $y_1 = 1, y_2 = 2$ e $y_3 = 5$. A soma dos produtos é:

$$\sum_{i=1}^{3} x_i y_i = 2 \cdot 1 + 3 \cdot 2 + 0 \cdot 5 = 8$$

Para simplificar, muitas vezes se escreve $\Sigma x, \Sigma y, \Sigma xy, \Sigma x^2, (\Sigma x)^2, \Sigma y^2, (\Sigma y)^2$. Isto pode ser feito, desde que esteja claro quais são os números que devem ser somados.

15.2 - Análise Combinatória

Se n é um número inteiro positivo maior do que zero, por definição, *fatorial de n*, que se indica por $n!$, é

$$n! = n\ (n-1)\ (n-2)...1.$$

O fatorial de 5 é

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120.$$

O desenvolvimento de um fatorial pode ser interrompido antes de chegar ao número 1, desde que se coloque o símbolo !, que indica o fatorial, logo após o último número. Então, pode-se escrever:

$$5! = 5 \cdot 4 \cdot 3!$$

porque

$$3! = 3 \cdot 2 \cdot 1.$$

Então

$$5! = 5 \cdot 4 \cdot 3! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1.$$

O fatorial de zero, que se indica por 0! é, por definição, igual a 1.

Dado um conjunto de n elementos, onde $n \neq 0$, e dado o número $x \leq n$, *combinação de n, x a x*, que se indica por $\binom{n}{x}$ é:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Pode-se provar que esta fórmula dá o número de diferentes conjuntos de x elementos que podem ser formados com n elementos distintos. Dois conjuntos só serão diferentes se tiverem, pelo menos, um elemento distinto.

Seja $n = 5$ e $x = 3$. Então a combinação de 5, 3 a 3 é:

$$\binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{5 \cdot 4 \cdot 3!}{3!2!} = 10$$

Convém observar que, para todo n:

$$\binom{n}{1} = n,$$

$$\binom{n}{n} = 1,$$

$$\binom{n}{0} = 1.$$

15.3 - Equação da Reta

No sistema de eixos cartesianos, a equação

$$Y = a + bX$$

representa uma reta. O valor a é o *coeficiente linear da reta*, e o valor b é o *coeficiente angular da reta*.

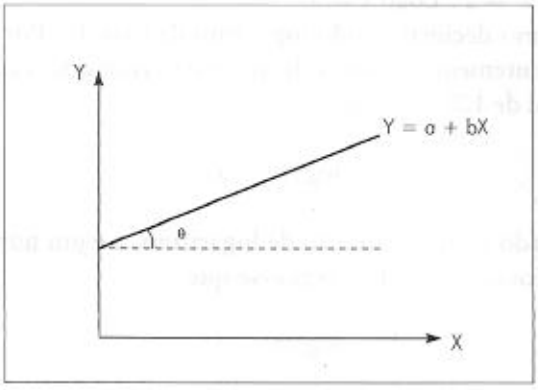
Como mostra a Figura 15.1, o valor de a dá a altura em que a reta corta o eixo das ordenadas. Então, se a é positivo, a reta corta o eixo das ordenadas acima da origem; se a é negativo, a reta corta o eixo das ordenadas abaixo da origem.

O valor b é a tangente trigonométrica do ângulo θ formado pelo eixo das abscissas e pela reta de equação $Y = a + bX$. Então b dá a inclinação da reta. Se o valor de b for positivo, a reta é ascendente, e se o valor de b for negativo, a reta é descendente.

Como exemplo, considere a reta de equação

$$Y = 8 - 2X$$

Figura 15.1 *Apresentação gráfica de uma reta* $Y = a + bX$

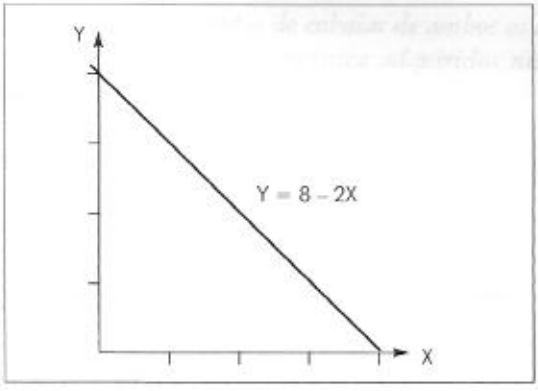


Para $X = 0$, $Y = 8$, e para $X = 3$,

$$Y = 8 - 2 \cdot 3 = 8 - 6 = 2.$$

Os pontos (0,8) e (3,2) pertencem à reta. Colocando esses pontos no sistema de eixos cartesianos, obtém-se a reta, mostrada na Figura 15.2.

Figura 15.2 *Apresentação gráfica da reta* $Y = 8 - 2X$



15.4 - LOGARITMOS

Sejam a e b dois números tais que $b > 0$ e $0 < a \neq 1$. Por definição, logaritmo de b na base a é um número x tal que $a^x = b$. Escreve-se

$$log_a^b = x$$

Para melhor entender a definição de logaritmo, considere que você quer obter o logaritmo de 8 na base 2. Escreva

$$log_2^8 = x$$

Dada a definição de logaritmo, tem-se que $2^x = 8$. Como $8 = 2^3$, segue-se que $2^x = 2^3$. Logo $x = 3$.

Logaritmo decimal é todo logaritmo de base 10. Por facilidade, a base é, freqüentemente, omitida. Imagine que você quer calcular o logaritmo decimal de 100. Escreva

$$\log 100 = x.$$

De acordo com a definição de logaritmo, x é um número tal que $10^x = 100$. Como $100 = 10^2$, segue-se que

$$\log 100 = 2.$$

Logaritmo neperiano é todo logaritmo de base e, um número irracional que vale 2,71828.... Indica-se logaritmo neperiano por ln. Os logaritmos neperianos são muito usados na teoria de Estatística.

Exercícios de Revisão

As técnicas estatísticas mais usuais na área de saúde foram apresentadas neste livro. Para dar idéia de como se aplicam essas técnicas em pesquisa, são fornecidos aqui alguns exemplos, todos baseados no trabalho de HOSSNE *et alii* (1990).

1 - ***São dados pesos e comprimentos de cobaias de ambos os sexos, com 25 dias. Usando os conhecimentos de estatística adquiridos neste livro, que análise você faria?***

Peso, em gramas, e comprimento, em centímetros, de cobaias de 25 dias segundo o sexo

Machos		Fêmeas	
Peso	Comprimento	Peso	Comprimento
200	20	225	20
200	19	225	20
215	22	200	19
215	21	200	18
215	20	200	19
155	19	200	19
160	18	200	18
285	19	200	19
190	19	200	19
230	19	200	18

Fonte: HOSSNE et alii (1990)

2 - *São dados os pesos frescos dos órgãos de cobaias machos com 90 dias de idade. Usando os conhecimentos de estatística adquiridos neste livro, que análise você faria?*

Peso fresco, em gramas, de órgãos de cobaias machos com 90 dias

Órgão		
Cérebro	Coração	Pulmões
2,77	1,61	3,14
2,79	1,65	3,63
2,52	1,98	3,63
2,84	1,46	4,27
2,48	1,54	3,44
2,72	1,44	3,74
2,76	1,41	3,29
2,52	1,38	3,72
2,35	1,67	3,19
2,41	1,33	3,27

Fonte: HOSSNE et alii (1990)

3 - *São dados os pesos frescos (PF) e os pesos secos (PS) de cérebros de cobaias de 90 dias. Usando os conhecimentos de estatística adquiridos neste livro, que análise você faria? Defina a quantidade (PF – PS)/PF, que dá a proporção de água no órgão. Multiplique por 100, para ter o resultado em porcentagem. Que análise estatística você faria desses resultados?*

Peso fresco e peso seco de cérebros de cobaias com 90 dias

Peso fresco	Peso seco
2,77	0,533
2,79	0,866
2,52	0,508
2,84	0,573
2,48	0,500
2,72	0,846
2,76	0,554
2,52	0,506
2,35	0,480
2,41	0,500

Fonte: HOSSNE et alii (1990)

4 - *São dados o tamanho da amostra (n), a média (x) e o desvio padrão (s) de pesos de cobaias de 25, 60 e 90 dias, machos e fêmeas. Que análise você faria?*

Tamanho da amostra, média e desvios padrões de pesos de cobaias, segundo a idade e o sexo

Idade	Machos			Fêmeas		
	n	*x*	*s*	*n*	*x*	*s*
25	10	196,5	24,5	10	205,0	10,5
60	10	310,0	31,6	10	313,0	16,5
90	10	500,0	10,0	10	502,0	6,3

Fonte: HOSSNE et alii (1990)

Apêndices

Tabelas

Tabela A.1

Distribuição normal reduzida $P(0 < Z < z)$

	Último dígito									
	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2703	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4658	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Tabela A.2

Valores de x^2, segundo os graus de liberdade e o valor de α

Graus de liberdade	α 10%	5%	1%
1	2,71	3,84	6,64
2	4,60	5,99	9,21
3	6,25	7,82	11,34
4	7,78	9,49	13,28
5	9,24	11,07	15,09
6	10,64	12,59	16,81
7	12,02	14,07	18,48
8	13,36	15,51	20,09
9	14,68	16,92	21,67
10	15,99	18,31	23,21
11	17,28	19,68	24,72
12	18,55	21,03	26,22
13	19,81	22,36	27,69
14	21,06	23,68	29,14
15	22,31	25,00	30,58
16	23,54	26,30	32,00
17	24,77	27,59	33,41
18	25,99	28,87	34,80
19	27,20	30,14	36,19
20	28,41	31,41	37,57
21	29,62	32,67	38,93
22	30,81	33,92	40,29
23	32,01	35,17	41,64
24	33,20	36,42	42,98
25	34,38	37,65	44,31
26	35,56	38,88	45,64
27	36,74	40,11	46,96
28	37,92	41,34	48,28
29	39,09	42,56	49,59
30	40,26	43,77	50,89

Tabela A.3

Valores de F para $\alpha = 2{,}5\%$, segundo o número de graus de liberdade do numerador e do denominador

Nº de g. l. do denominador	Número de graus de liberdade do numerador								
	1	2	3	4	5	6	7	8	9
1	648	800	864	900	922	937	948	957	963
2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4
3	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,5
4	12,2	10,6	9,98	9,60	9,36	9,20	9,07	8,98	8,90
5	10,0	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11

continua

Continuação da Tabela A.3

Nº de g. l. do denominador	Número de graus de liberdade do numerador									
	10	12	15	20	24	30	40	60	120	∞
1	969	977	985	993	997	1000	1010	1010	1010	1020
2	39,4	39,4	39,4	39,4	39,5	39,5	39,5	39,5	39,5	39,5
3	14,4	14,3	14,3	14,2	14,1	14,1	14,0	14,0	13,9	13,9
4	8,84	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26
5	6,62	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07	6,02
6	5,46	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
7	4,76	4,67	4,57	4,47	4,42	4,36	4,31	4,25	4,20	4,14
8	4,30	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67
9	3,96	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
10	3,72	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08
11	3,53	3,43	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88
12	3,37	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,72
13	3,25	3,15	3,05	2,95	2,89	2,84	2,78	2,72	2,66	2,60
14	3,15	3,05	2,95	2,84	2,79	2,73	2,67	2,61	2,55	2,49
15	3,06	2,96	2,86	2,76	2,70	2,64	2,59	2,52	2,46	2,40
16	2,99	2,89	2,79	2,68	2,63	2,57	2,51	2,45	2,38	2,32
17	2,92	2,82	2,72	2,62	2,56	2,50	2,44	2,38	2,32	2,25
18	2,87	2,77	2,67	2,56	2,50	2,44	2,38	2,32	2,26	2,19
19	2,82	2,72	2,62	2,51	2,45	2,39	2,33	2,27	2,20	2,13
20	2,77	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09
21	2,73	2,64	2,53	2,42	2,37	2,31	2,25	2,18	2,11	2,04
22	2,70	2,60	2,50	2,39	2,33	2,27	2,21	2,14	2,08	2,00
23	2,67	2,57	2,47	2,36	2,30	2,24	2,18	2,11	2,04	1,97
24	2,64	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94
25	2,61	2,51	2,41	2,30	2,24	2,18	2,12	2,05	1,98	1,91
26	2,59	2,49	2,39	2,28	2,22	2,16	2,09	2,03	1,95	1,88
27	2,57	2,47	2,36	2,25	2,19	2,13	2,07	2,00	1,93	1,85
28	2,55	2,45	2,34	2,23	2,17	2,11	2,05	1,98	1,91	1,83
29	2,53	2,43	2,32	2,21	2,15	2,09	2,03	1,96	1,89	1,81
30	2,51	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79
40	2,39	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64
60	2,27	2,17	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48
120	2,16	2,05	1,94	1,82	1,76	1,69	1,61	1,53	1,43	1,31
∞	2,05	1,94	1,83	1,71	1,64	1,57	1,48	1,39	1,27	1,00

Fonte: SCHEFFÉ (1959)

Tabela A.4

Valores de F para $\alpha = 5\%$, segundo o número de graus de liberdade do numerador e do denominador

Nº de g. l. do denominador	Número de graus de liberdade do numerador								
	1	2	3	4	5	6	7	8	9
1	161	200	216	225	230	234	237	239	241
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88

Continuação da Tabela A.4

Nº de g. l. do denominador	Número de graus de liberdade do numerador									
	10	12	15	20	24	30	40	60	120	∞
1	242	244	246	248	249	250	251	252	253	254
2	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5
3	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Fonte: SCHEFFÉ (1959)

Tabela A.5

Valores de F para $\alpha = 10\%$, segundo o número de graus de liberdade do numerador e do denominador

Nº de g. l. do denominador	Número de graus de liberdade do numerador								
	1	2	3	4	5	6	7	8	9
1	39,9	49,5	53,6	55,8	57,2	58,2	58,9	59,4	59,9
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96
21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93
23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88
27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87
29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79
60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74
120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68
∞	2,71	2,30	2,08	1,94	1,85	1,77	1,72	1,67	1,63

continua

Continuação da Tabela A.5

Nº de g. l. do denominador	Número de graus de liberdade do numerador									
	10	12	15	20	24	30	40	60	120	∞
1	60,2	60,7	61,2	61,7	62,0	62,3	62,5	62,8	63,1	63,3
2	9,39	9,41	9,42	9,44	9,45	9,46	9,47	9,47	9,48	9,49
3	5,23	5,22	5,20	5,18	5,18	5,17	5,16	5,15	5,14	5,13
4	3,92	3,90	3,87	3,84	3,83	3,82	3,80	3,79	3,78	3,76
5	3,30	3,27	3,24	3,21	3,19	3,17	3,16	3,14	3,12	3,10
6	2,94	2,90	2,87	2,84	2,82	2,80	2,78	2,76	2,74	2,72
7	2,70	2,67	2,63	2,59	2,58	2,56	2,54	2,51	2,49	2,47
8	2,54	2,50	2,46	2,42	2,40	2,38	2,36	2,34	2,32	2,29
9	2,42	2,38	2,34	2,30	2,28	2,25	2,23	2,21	2,18	2,16
10	2,32	2,28	2,24	2,20	2,18	2,16	2,13	2,11	2,08	2,06
11	2,25	2,21	2,17	2,12	2,10	2,08	2,05	2,03	2,00	1,97
12	2,19	2,15	2,10	2,06	2,04	2,01	1,99	1,96	1,93	1,90
13	2,14	2,10	2,05	2,01	1,98	1,96	1,93	1,90	1,88	1,85
14	2,10	2,05	2,01	1,96	1,94	1,91	1,89	1,86	1,83	1,80
15	2,06	2,02	1,97	1,92	1,90	1,87	1,85	1,82	1,79	1,76
16	2,03	1,99	1,94	1,89	1,87	1,84	1,81	1,78	1,75	1,72
17	2,00	1,96	1,91	1,86	1,84	1,81	1,78	1,75	1,72	1,69
18	1,98	1,93	1,89	1,84	1,81	1,78	1,75	1,72	1,69	1,66
19	1,96	1,91	1,86	1,81	1,79	1,76	1,73	1,70	1,67	1,63
20	1,94	1,89	1,84	1,79	1,77	1,74	1,71	1,68	1,64	1,61
21	1,92	1,88	1,83	1,78	1,75	1,72	1,69	1,66	1,62	1,59
22	1,90	1,86	1,81	1,76	1,73	1,70	1,67	1,64	1,60	1,57
23	1,89	1,84	1,80	1,74	1,72	1,69	1,66	1,62	1,59	1,55
24	1,88	1,83	1,78	1,73	1,70	1,67	1,64	1,61	1,57	1,53
25	1,87	1,82	1,77	1,72	1,69	1,66	1,63	1,59	1,56	1,52
26	1,86	1,81	1,76	1,71	1,68	1,65	1,61	1,58	1,54	1,50
27	1,85	1,80	1,75	1,70	1,67	1,64	1,60	1,57	1,53	1,49
28	1,84	1,79	1,74	1,69	1,66	1,63	1,59	1,56	1,52	1,48
29	1,83	1,78	1,73	1,68	1,65	1,62	1,58	1,55	1,51	1,47
30	1,82	1,77	1,72	1,67	1,64	1,61	1,57	1,54	1,50	1,46
40	1,76	1,71	1,66	1,61	1,57	1,54	1,51	1,47	1,42	1,38
60	1,71	1,66	1,60	1,54	1,51	1,48	1,44	1,40	1,35	1,29
120	1,65	1,60	1,55	1,48	1,45	1,41	1,37	1,32	1,26	1,19
∞	1,60	1,55	1,49	1,42	1,38	1,34	1,30	1,24	1,17	1,00

Fonte: SCHEFFÉ (1959)

Tabela A.6

Valores de t, segundo os graus de liberdade e o valor de α

Graus de liberdade	α 10%	5%	1%
1	6,31	12,71	63,66
2	2,92	4,30	9,92
3	2,35	3,18	5,84
4	2,13	2,78	4,60
5	2,02	2,57	4,03
6	1,94	2,45	3,71
7	1,90	2,36	3,50
8	1,86	2,31	3,36
9	1,83	2,26	3,25
10	1,81	2,23	3,17
11	1,80	2,20	3,11
12	1,78	2,18	3,06
13	1,77	2,16	3,01
14	1,76	2,14	2,98
15	1,75	2,13	2,95
16	1,75	2,12	2,92
17	1,74	2,11	2,90
18	1,73	2,10	2,88
19	1,73	2,09	2,86
20	1,73	2,09	2,84
21	1,72	2,08	2,83
22	1,72	2,07	2,82
23	1,71	2,07	2,81
24	1,71	2,06	2,80
25	1,71	2,06	2,79
26	1,71	2,06	2,78
27	1,70	2,05	2,77
28	1,70	2,05	2,76
29	1,70	2,04	2,76
30	1,70	2,04	2,75
40	1,68	2,02	2,70
60	1,67	2,00	2,66
120	1,66	1,98	2,62
∞	1,64	1,96	2,58

Tabela A.7

Valores da amplitude total estudentizada (q) para $\alpha = 5\%$, segundo o número de tratamento (k) e os graus de liberdade do resíduo

Nº de graus de lib. do resíduo	Número de tratamentos (k)																		
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	8,0	27,0	32,8	37,1	40,4	43,1	45,4	47,4	49,1	50,6	52,0	53,2	54,3	55,4	56,3	57,2	58,0	58,8	59,6
2	6,08	8,33	9,80	10,9	11,7	12,4	13,0	13,5	14,0	14,4	14,7	15,1	15,4	15,7	15,9	16,1	16,4	16,6	16,8
3	4,50	5,91	6,82	7,50	8,04	8,48	8,85	9,18	9,46	9,72	9,95	10,2	10,3	10,5	10,7	10,8	11,0	11,1	11,2
4	3,93	5,04	5,76	6,29	6,71	7,05	7,35	7,60	7,83	8,03	8,21	8,37	8,52	8,66	8,79	8,91	9,03	9,13	9,23
5	3,64	4,60	5,22	5,67	6,03	6,33	6,58	6,80	6,99	7,17	7,32	7,47	7,60	7,72	7,83	7,93	8,03	8,12	8,21
6	3,46	4,34	4,90	5,30	5,63	5,90	6,12	6,32	6,49	6,65	6,79	6,92	7,03	7,14	7,24	7,34	7,43	7,51	7,59
7	3,34	4,16	4,68	5,06	5,36	5,61	5,82	6,00	6,16	6,30	6,43	6,55	6,66	6,76	6,85	6,94	7,02	7,10	7,17
8	3,26	4,04	4,53	4,89	5,17	5,40	5,60	5,77	5,92	6,05	6,18	6,29	6,39	6,48	6,57	6,65	6,73	6,80	6,87
9	3,20	3,95	4,41	4,76	5,02	5,24	5,43	5,59	5,74	5,87	5,98	6,09	6,19	6,28	6,36	6,44	6,51	6,58	6,64
10	3,15	3,88	4,33	4,65	4,91	5,12	5,30	5,46	5,60	5,72	5,83	5,93	6,03	6,11	6,19	6,27	6,34	6,40	6,47
11	3,11	3,82	4,26	4,57	4,82	5,03	5,20	5,35	5,49	5,61	5,71	5,81	5,90	5,98	6,06	6,13	6,20	6,27	6,33
12	3,08	3,77	4,20	4,51	4,75	4,95	5,12	5,27	5,39	5,51	5,61	5,71	5,80	5,88	5,95	6,02	6,09	6,15	6,21
13	3,06	3,73	4,15	4,45	4,69	4,88	5,05	5,19	5,32	5,43	5,53	5,63	5,71	5,79	5,86	5,93	5,99	6,05	6,11
14	3,03	3,70	4,11	4,41	4,64	4,83	4,99	5,13	5,25	5,36	5,46	5,55	5,64	5,71	5,79	5,85	5,91	5,97	6,03
15	3,01	3,67	4,08	4,37	4,59	4,78	4,94	5,08	5,20	5,31	5,40	5,49	5,57	5,65	5,72	5,78	5,85	5,90	5,96
16	3,00	3,65	4,05	4,33	4,56	4,74	4,90	5,03	5,15	5,26	5,35	5,44	5,52	5,59	5,66	5,73	5,79	5,84	5,90
17	2,98	3,63	4,02	4,30	4,52	4,70	4,86	4,99	5,11	5,21	5,31	5,39	5,47	5,54	5,61	5,67	5,73	5,79	5,84
18	2,97	3,61	4,00	4,28	4,49	4,67	4,82	4,96	5,07	5,17	5,27	5,35	5,43	5,50	5,57	5,63	5,69	5,74	5,79
19	2,96	3,59	3,98	4,25	4,47	4,65	4,79	4,92	5,04	5,14	5,23	5,31	5,39	5,46	5,53	5,59	5,65	5,70	5,75
20	2,95	3,58	3,96	4,23	4,45	4,62	4,77	4,90	5,01	5,11	5,20	5,28	5,36	5,43	5,49	5,55	5,61	5,66	5,71
24	2,92	3,53	3,90	4,17	4,37	4,54	4,68	4,81	4,92	5,01	5,10	5,18	5,25	5,32	5,38	5,44	5,49	5,55	5,59
30	2,89	3,49	3,85	4,10	4,30	4,46	4,60	4,72	4,82	4,92	5,00	5,08	5,15	5,21	5,27	5,33	5,38	5,43	5,47
40	2,86	3,44	3,79	4,04	4,23	4,39	4,52	4,63	4,73	4,82	4,90	4,98	5,04	5,11	5,16	5,22	5,27	5,31	5,36
60	2,83	3,40	3,74	3,98	4,16	4,31	4,44	4,55	4,65	4,73	4,81	4,88	4,94	5,00	5,06	5,11	5,15	5,20	5,24
120	2,80	3,36	3,68	3,92	4,10	4,24	4,36	4,47	4,56	4,64	4,71	4,78	4,84	4,90	4,95	5,00	5,04	5,09	5,13
∞	2,77	3,31	3,63	3,86	4,03	4,17	4,29	4,39	4,47	4,55	4,62	4,68	4,74	4,80	4,85	4,89	4,93	4,97	5,01

Fonte: SCHEFFÉ, (1959)

Tabela A.8

Valores da amplitude total estudentizada (q) para $\alpha = 10\%$, segundo o número de tratamento (k) e os graus de liberdade do resíduo

Nº de graus de lib. do resíduo	Número de tratamentos (k)																		
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	8,93	13,4	16,4	18,5	20,2	21,5	22,6	23,6	24,5	25,2	25,9	26,5	27,1	27,6	28,1	28,5	29,0	29,3	29,7
2	4,13	5,73	6,77	7,54	8,14	8,63	9,05	9,41	9,72	10,0	10,3	10,5	10,7	10,9	11,1	11,2	11,4	11,5	11,7
3	3,33	4,47	5,20	5,74	6,16	6,51	6,81	7,06	7,29	7,49	7,67	7,83	7,98	8,12	8,25	8,37	8,48	8,58	8,68
4	3,01	3,98	4,59	5,03	5,39	5,68	5,93	6,14	6,33	6,49	6,65	6,78	6,91	7,02	7,13	7,23	7,33	7,41	7,50
5	2,85	3,72	4,26	4,66	4,98	5,24	5,46	5,65	5,82	5,97	6,10	6,22	6,34	6,44	6,54	6,63	6,71	6,79	6,86
6	2,75	3,56	4,07	4,44	4,73	4,97	5,17	5,34	5,50	5,64	5,76	5,87	5,98	6,07	6,16	6,25	6,32	6,40	6,47
7	2,68	3,45	3,93	4,28	4,55	4,78	4,97	5,14	5,28	5,41	5,53	5,64	5,74	5,83	5,91	5,99	6,06	6,13	6,19
8	2,63	3,37	3,83	4,17	4,43	4,65	4,83	4,99	5,13	5,25	5,36	5,46	5,56	5,64	5,72	5,80	5,87	5,93	6,00
9	2,59	3,32	3,76	4,08	4,34	4,54	4,72	4,87	5,01	5,13	5,23	5,33	5,42	5,51	5,58	5,66	5,72	5,79	5,85
10	2,56	3,27	3,70	4,02	4,26	4,47	4,64	4,78	4,91	5,03	5,13	5,23	5,32	5,40	5,47	5,54	5,61	5,67	5,73
11	2,54	3,23	3,66	3,96	4,20	4,40	4,57	4,71	4,84	4,95	5,05	5,15	5,23	5,31	5,38	5,45	5,51	5,57	5,63
12	2,52	3,20	3,62	3,92	4,16	4,35	4,51	4,65	4,78	4,89	4,99	5,08	5,16	5,24	5,31	5,37	5,44	5,49	5,55
13	2,50	3,18	3,59	3,88	4,12	4,30	4,46	4,60	4,72	4,83	4,93	5,02	5,10	5,18	5,25	5,31	5,37	5,43	5,48
14	2,49	3,16	3,56	3,85	4,08	4,27	4,42	4,56	4,68	4,79	4,88	4,97	5,05	5,12	5,19	5,26	5,32	5,37	5,43
15	2,48	3,14	3,54	3,83	4,05	4,23	4,39	4,52	4,64	4,75	4,84	4,93	5,01	5,08	5,15	5,21	5,27	5,32	5,38
16	2,47	3,12	3,52	3,80	4,03	4,21	4,36	4,49	4,61	4,71	4,81	4,89	4,97	5,04	5,11	5,17	5,23	5,28	5,33
17	2,46	3,11	3,50	3,78	4,00	4,18	4,33	4,46	4,58	4,68	4,77	4,86	4,93	5,01	5,07	5,13	5,19	5,24	5,30
18	2,45	3,10	3,49	3,77	3,98	4,16	4,31	4,44	4,55	4,65	4,75	4,83	4,90	4,98	5,04	5,10	5,16	5,21	5,26
19	2,45	3,09	3,47	3,75	3,97	4,14	4,29	4,42	4,53	4,63	4,72	4,80	4,88	4,95	5,01	5,07	5,13	5,18	5,23
20	2,44	3,08	3,46	3,74	3,95	4,12	4,27	4,40	4,51	4,61	4,70	4,78	4,85	4,92	4,99	5,05	5,10	5,16	5,20
24	2,42	3,05	3,42	3,69	3,90	4,07	4,21	4,34	4,44	4,54	4,63	4,71	4,78	4,85	4,91	4,97	5,02	5,07	5,12
30	2,40	3,02	3,39	3,65	3,85	4,02	4,16	4,28	4,38	4,47	4,56	4,64	4,71	4,77	4,83	4,89	4,94	4,99	5,03
40	2,38	2,99	3,35	3,60	3,80	3,96	4,10	4,21	4,32	4,41	4,49	4,56	4,63	4,69	4,75	4,81	4,86	4,90	4,95
60	2,36	2,96	3,31	3,56	3,75	3,91	4,04	4,16	4,25	4,34	4,42	4,49	4,56	4,62	4,67	4,73	4,78	4,82	4,86
120	2,34	2,93	3,28	3,52	3,71	3,86	3,99	4,10	4,19	4,28	4,35	4,42	4,48	4,54	4,60	4,65	4,69	4,74	4,78
	2,33	2,90	3,24	3,48	3,66	3,81	3,93	4,04	4,13	4,21	4,28	4,35	4,41	4,47	4,52	4,57	4,61	4,65	4,69

Fonte: SCHEFFÉ, (1959)

Capítulo 1

1.6.1 - Podem ser obtidas 6 amostras diferentes:
Antônio e Luís; Antônio e Pedro; Antônio e Carlos; Luís e Pedro; Luís e Carlos; Pedro e Carlos.

1.6.2 - Podem ser selecionados: a) os elementos de ordem par; b) os elementos de ordem ímpar; c) os 4 primeiros elementos.

1.6.3 - Numeram-se os alunos e sorteiam-se seis.

1.6.5 - O tipo de serviço odontológico que uma família demanda depende da sua renda. A amostragem com base na lista telefônica é incorreta porque seleciona apenas aqueles suficientemente ricos para ter um telefone.

Capítulo 2

2.5.1 - A tabela apresentada em seguida mostra que os pedestres são, proporcionalmente, as maiores vítimas fatais de acidentes de trânsito.

Vítimas fatais de acidentes de trânsito no Brasil em 1986

Tipo	Freqüência	Percentual
Pedestres	11 712	42,89
Passageiros	7 116	26,06
Condutores	8 478	31,05
Total	27 306	100,00

Fonte: IBGE (1988)

2.5.2 - A tabela apresentada em seguida sugere que o risco de morte após três anos provavelmente não depende da faixa de idade por ocasião do diagnóstico de câncer de mama.

Pacientes com câncer de mama segundo a faixa de idade por ocasião do diagnóstico e a sobrevivência após três anos

Faixa de idade	Sobrevivência		Freqüência relativa de não-sobreviventes
	Sim	Não	
Menos de 50 anos	11	6	35,3
De 50 a 70 anos	18	8	30,8
Mais de 70 anos	15	9	37,5

2.5.3 - A tabela é:

Estabelecimentos de saúde públicos e particulares por espécie. Brasil, 1986

Espécie	Estabelecimentos	
	Públicos	Particulares
Hospital	16,3	83,7
Pronto-socorro	49,0	51,0
Policlínicas	20,0	80,0
Outros (1)	96,8	3,2

Fonte: IBGE (1988)
(1) - Inclui postos de saúde, centros de saúde e unidades mistas.

2.5.4 - Usando intervalos iguais, tem-se:

Distribuição do tempo de internação, em dias, de pacientes acidentados no trabalho, em um dado hospital

Classe	Freqüência
0 ⊢ 2	2
2 ⊢ 4	6
4 ⊢ 6	5
6 ⊢ 8	7
8 ⊢ 10	5
10 ⊢ 12	3
12 ⊢ 14	4
14 ⊢ 16	3
16 ⊢ 18	1
Total	36

Usando intervalos diferentes, tem-se:

Distribuição do tempo de internação, em dias, de pacientes acidentados no trabalho, em um dado hospital

Classe	Freqüência
1 dia	2
2 ou 3 dias	6
De 4 a 7 dias	12
De 8 a 14 dias	14
Mais de 14 dias	2
Total	36

Capítulo 4

4.6.1 - A média é 90,625mg.

4.6.2 - A mediana é 83g.

4.6.3 - As médias são 0,9 dente cariado para meninos e 1,0 dente cariado para meninas.

4.6.4 - A moda é doença mental, ou seja, esta é a causa mais freqüentemente atribuída ao suicídio.

Capítulo 5

5.6.1 - Tem-se que $s^2 = 79{,}911$, $s^2 = 8{,}939$ e CV = 9,86%.

5.6.2 - A amplitude é 15g.

5.5.3 - Tem-se, para machos, $\bar{x} = 25{,}45$ e $s = 0{,}725$ e, para fêmeas, $\bar{x} = 26{,}95$ e $s = 0{,}599$.

5.6.4 - Tem-se, para pulmão direito, $\bar{x} = 2{,}02$ e $s = 0{,}26$ e, para pulmão esquerdo, $\bar{x} = 1{,}62$ e $s = 0{,}17$.

Capítulo 6

6.5.1 - $r = 0$, não existe correlação.

6.5.2 - $\Sigma x = 41{,}80$, $\Sigma x^2 = 293{,}16$, $\Sigma y = 12{,}60$, $\Sigma y^2 = 26{,}58$, $\Sigma xy = 88{,}23$. Logo, $r = 0{,}9295$.

6.5.3 - $\Sigma x = 255$, $\Sigma x^2 = 9\,443$, $\Sigma y = 17{,}25$, $\Sigma y^2 = 50{,}4375$, $\Sigma xy = 660{,}25$. Logo, $r = 0{,}9125$.

6.5.4 - A simples observação dos dados sugere que existe correlação positiva entre o índice clínico e o peso seco das placas dentais, porque os dados crescem juntos.

Capítulo 7

7.6.1 - O gráfico de linhas mostra que o peso médio dos 8 ratos aumentou com a idade, no período observado.
7.6.2 - $a = 4, b = 0$. Então $\hat{Y} = 4$; a reta é paralela ao eixo das abscissas.
7.6.3 - $\hat{Y} = 0{,}495 + 0{,}2304X$
7.6.4 - $\hat{Y} = -5{,}09 + 0{,}207X$.

Capítulo 8

8.7.1 - a) 50% b) 50%
8.7.2 - 0,1%
8.7.3 - 50%
8.7.4 - a) 36% b) 1%
8.7.5 - 50%

Capítulo 9

9.6.1 - A tabela fica como segue:

Distribuição do número de meninos em uma família de 5 crianças

X	P(X)
0	1/32
1	5/32
2	10/32
3	10/32
4	5/32
5	1/32

9.6.2 - $\mu = 5, \sigma^2 = 2{,}5$
9.6.3 - $\mu = 2, \sigma^2 = 1{,}6$
9.6.4 - 2,7%
9.6.5 - 27/64 ou 42,2%
9.6.6 - 0,001%

Capítulo 10

10.6.1 - a) 78,88% b) 10,56%
10.6.2 - a) 4,75% b) 45,25%

10.6.3 - a) 97,72% b) 2,28%
10.6.4 - a) 21,19% b) 21,19%

Capítulo 11

11.9.1 - $\chi^2 = 4{,}82$. A proporção de recém-nascidos portadores de anomalia congênita é maior no sexo feminino.
11.9.2 - $\chi^2 = 0{,}65$. A proporção de pessoas Rh⁻ não depende da origem.
11.9.3 - $\chi^2 = 1{,}32$. A presença de aberração cromossômica no feto não depende da idade da gestante.
11.9.4 - $\chi^2 = 9{,}04$. A ausência congênita de dentes ocorre mais em meninas.

Capítulo 12

12.9.1 - A tabela dada em seguida apresenta médias e desvios padrões de pesos de ratos.

Médias e desvios padrões de pesos de ratos

Ração	$\bar{x}$	s
Padrão	188,0	3,7
Experimental	212,0	3,7

O valor de t é 4,536, significante a 5%. Os ratos submetidos à ração experimental ganharam mais peso.

12.9.2 - Observações pareadas; $t = 4{,}226$, significante a 5%. O teste B dá maior resultado de QI.
12.9.3 - $t = 1{,}642$, não-significante a 5%. Os dados não mostram que o uso de anovulatório aumenta a pressão sangüínea sistólica.
12.9.4 - $t = 0{,}623$, não-significante a 5%. Os dados não mostram diferença de peso ao nascer entre sexos.

Capítulo 13

13.5.1 - A tabela de análise de variância deve ser apresentada como segue:

Análise de variância

Causas de variação	GL	SQ	QM	F
Tratamentos	4	184,00	46,00	4,60
Resíduo	15	150,00	10,00	
Total	19	334,00		

O valor de F é significante ao nível de 5%. As médias são, respectivamente: 17; 11; 19; 12; 16. A d.m.s. é 6,91. A média do tratamento C é significativamente maior do que as médias dos tratamentos B e D.

13.5.2 - A tabela de análise de variância deve ser apresentada como segue:

Análise de variância

Causas de variação	GL	SQ	QM	F
Tratamentos	3	308,00	102,667	5,70
Resíduo	11	198,00	18,000	
Total	14	506,00		

O valor de F é significante ao nível de 5%. As médias de tratamentos são, respectivamente: 16; 5; 15; 9. O teste de Tukey mostra que as médias dos tratamentos A e C são significativamente maiores do que a média de B.

Capítulo 14

14.5.1 - $121{,}7 < \mu < 124{,}3$
14.5.2 - $15{,}50 < \mu < 16{,}90$
14.5.3 - $49{,}20 < \mu < 50{,}80$
14.5.4 - $92{,}5 < \mu < 97{,}5$